APPLICATION OF NON-LINEAR REGRESSION TO CAR OWNERSHIP IN JOHANNESBURG

John Kelly

Transportation Information and Technology sub-DirectorateCity of Johannesburg, P.O. Box 30733,
Braamfontein 2017
johnke@joburg.org.za

ABSTRACT

An optimal smoothing algorithm is developed and applied to Moving SA (Johannesburg) data. Non-Linear regression is applied successfully to the Moving SA data. Values of 437.4 for saturation (vehicles/k.capita), 0.04569 for growth parameter (per annum frequency), and 1979.77 (years) for point of maximum growth are obtained. A correlation of +0.9647 with the raw data was achieved. The soundness of the statistical inference is demonstrated by showing the bias and trend of the model error term to be statistically indistinguishable from zero. Inner bounds on the confidence intervals for small sample non-linear regression are constructed for the predicted values.

INTRODUCTION

Motivation

Practical Applications

There are numerous possible applications of the car ownership logistic model, some of which are listed below. The model can be used to: -

1. evaluate the feasibility of plans and policies with respect to their meeting demand for road space;
2. size and cost projects;
3. schedule the implementation of plans;
4. provide a base line prediction against which to monitor the effectiveness of interventions.
5. A further benefit of a good fit of the logistic modal is that the removal of the trend component of the time series is a necessary prerequisite for the analysis of the short and medium term oscillatory phenomena such as the business cycle.

Calibration

Calibration is a necessary step in the modelling process. The word calibration carries the connotation of a routine uncontroversial process. The analogy suggested is with calibration of an instrument of measurement. In calibrating an instrument for measuring a certain variable, a test specimen of known value of the variable is subject to measurement. The controls for the parameter of the instrument are then manipulated until the prior known reading for the variable is obtained. The ‘calibration’ process, in transportation modelling, is profoundly different. It usually involves the subjective removal of a significant number of data points.

Empiricism

A fundamental tenet of the scientific method is that data decide the validity of theories, and consequently theories should not decide the validity of data. However, in practice it often happens that some data are eliminated on the basis of *a priori* considerations. For example, it is not unusual for skilled and experienced modellers to discard as much as one third of the data in fitting the logistic DE, thereby seriously compromising the scientific validity of the model. This problem is not confined to the logistic model, but is widely found in the four-stage model.

Engineering Paradigm

How has this fundamental violation of empiricism crept into practice? Most problems addressed in engineering have been concerned with the application of Newtonian mechanics to designed macro-physical systems well within the classical domain. In this class of applications, Newtonian mechanics has been, with few exceptions, very successful. The very reasonable temptation, therefore, arises to treat data points substantially deviating from the theory (i.e. outliers) as errors.

Unfortunately, transportation planning is far removed from the comfort zone of Newtonian mechanics. Transportation planning is very much in the area of behavioural systems, especially socio-economic systems. Hence, much greater rigour is required. It can no longer be safely assumed that 'outliers' are errors. Serious consideration must be given to the possibility that these 'outliers' imply a systematic departure of reality from theory. That is, they constitute a disconfirmation of the accepted theory.

Moreover, socio-economic systems are complex, depend on a large number of variables and are interactive. The result is that there is a high noise level in the data, that is, the variance about the predicted curve is high and sometimes with systematic departures from the model.

Unlike experimentation on a physical system the variables are not easily manipulated. Hence, traditional experimental methods cannot be used.

Important Decisions

Moreover, substantial public and private investment decisions are informed by the results of transport models. Often, there is considerable political, commercial or professional interest vested in an outcome. As a consequence of the shortcomings of 'calibration', and the gravity of the outcomes, modellers are in a vulnerable position. If a valid and objective method could be found to handle the problems of 'calibration' not only could sounder and more accurate results be obtained, but modellers could reduce the risks arising from the 'calibration' process in possibly contentious circumstances.

The Dilemma and its Resolution

The dilemma faced by the practitioner is how to retain scientific objectivity while obtaining a good fit. A resolution of this dilemma is proposed and illustrated in this paper and in Kelly 2007. The solution proposed is to use statistical methods that retain all the data points, but reduce the influence, on the fit, of points depending on there deviation from the dominant trend. These methods are wholly objective, explicit and algorithmic.

The proposed method cannot find an explanation for the outliers, even less can the method construct a theory inclusive of the outliers. That it can do, most cases, is to find an objective best fit for existing theories. This is an improvement most practitioners would welcome.

Value of Greater Accuracy

A section of arterial road recently constructed in Soweto cost 23.1 M(illion)R/km (John White of JRA in private communication). The section of road was located in a suburban area and did not involve any bridges, or other major structures or any major cut or fill. The costs excluded land procurement, but included storm water reticulation, relocation of services and provision of street lighting. For the purposes of argument this cost can be taken as representative of marginal road construction costs (of building an extra km) in the City of Johannesburg (CoJ) area.

The Integrated Transport Plan 2003-2008 vol. 1 p. 39 table 3-11 states that there are 1,260 km of arterial roads on the CoJ. Given these data a low estimate of the approximate replacement cost of the arterial network is 29.1 Billion R.

The saturation parameter α can be taken as an indicator of long-term demand. The algorithm in Kelly 2007 gives $\hat{\alpha} = 484.7$ (the hat indicates that the value has been inferred from the data). The algorithm in this paper gives $\hat{\alpha} = 437.4$. Hence, the difference is $\Delta\hat{\alpha} = -47.3$ i.e. –10.8% of demand. This gives a 3.1 Billion R saving in previously planned expenditure. Such a saving should be sufficient reason to wish to improve the accuracy of the inferred values of the parameters.

Goals

1. To develop a minimum bias smoothing algorithm.
2. To develop a non-linear regression algorithm, which can successfully be applied to the car ownership data set out in the report Moving SA (Johannesburg).
3. To demonstrate the soundness of the statistical inference by showing that
   a. the biases on the parameters are statistically indistinguishable from zero,
   b. the bias and trend of the model error term (residual) is statistically indistinguishable from zero.
4. To construct confidence intervals for predicted values from a non-linear regression model applied to the Moving SA (small sample) data.
5. To fully automate the algorithm in S^+, a statistics package and programming language.

This paper does not attempt to build a modal quantifying the effect of variables causative of car ownership, but confines itself to time only as a surrogate of the causative variables. The role of the specific causative variables will be addressed in a later paper.

Cross-Pollination

Some may feel that this paper would be better placed in a statistics journal. With some adjustments and additions, it would be useful to publish in such a journal. Nonetheless, it is important to publish the work in a transportation context. The paper not only reports a promising algorithm for solving a problem that has troubled practitioners since logistic models were first applied to car ownership in the mid nineteen sixties. It also, attempts a cross-pollination between transport modelling and newer statistical methods such as non-linear regression, robust inference, and local regression. Moreover, these methods could find wider application in transportation modelling than vehicle ownership.

That many problems in the social sciences (of which transportation planning is one) reduce to a problem in statistics constitutes sufficient reason to place the paper in a transportation journal.

INVERSION PROBLEM FOR THE LOGISTIC DE

The Logistic Differential Equation

As discussed in Kelly 2007, the **logistic differential equation** (DE) is traditionally used to model the growth in car ownership. The logistic DE can be written as:

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

$$x(t_0) = x_0 \qquad (2),$$

where t is time. The time unit used in this study is the year, and 0h00 1st January in the year zero is $t = 0$. Car ownership level (ownership per thousand 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 an example of a state variable, and α and κ are referred to as parameters. A closed form solution for the above DE is: $x(t) = \alpha\left(1 + e^{-\kappa(t-\gamma)}\right)^{-1}$ ....(3). The parameter γ is the time shift parameter arising, as a constant of integration, out of the initial conditions of the DE (equation 2), and representing the point in time of maximum growth.

The problem of inferring the parameters α, κ, γ of the DE from a sample of observations (x_i, t_i) for $i = 1$ to n, constitutes what is known as an inverse problem for the DE.

Non-Linear Regression

In Kelly (2007 pp. 8-9), an attempt was made to apply non-linear regression to the inverse problem for the logistic differential equation as applied to car ownership. By simulation, the paper showed considerable gains in accuracy over the standard practice method of discretization of the derivative and a linerizing transformation. However, the algorithm supplied in S^+ failed to converge on the "Moving SA" (November 1997) data. This paper reports the development and application of a successful method of applying non-linear regression to the Moving SA data.

In the non-linear approach the closed form solution (e.g. 1) for the DE is used in the least squares method to infer the parameters values. Note the **regression is non-linear** because the model is non-linear in the parameters. It is possible to use the closed form solution, because the approximate value of the growth parameter ($\kappa \ll 3$), obtained from the linearizing method is safely below the chaotic region for the logistic DE.

As in the preceding paper (Kelly 2007) the algorithm used is a Gauss-Newton method, "nls" of S^+ ("S-Plus 2000 Guide to Statistics vol. 1" 1999). The Gauss-Newton method is an adaptation of the Newton method to least sum of squares problem devised by C.F. Gauss. Unlike the Newton algorithm the Gauss-Newton algorithm does not require us to compute the second derivatives of the squared errors.

The innovation is to apply smoothing to the data prior to the non-linear regression. The smoothing need not compromise the soundness of inference since the logistic DE model is a long-term trend model and the smoothing algorithm parameters are chosen to minimize bias.

OPTIMAL SMOOTHING

Purpose of Smoothing

Smoothing is not used in the usual way that a transformation is used in statistics. That is, it is not used to transform data with a non-Gaussian distribution into data with a Gaussian distributed data, so that classical statistical methods are applicable. Data smoothing does not attempt to fulfil this role.

The purpose of smoothing is noise reduction. The method of solution used to find the minimum error squared is the Gauss-Newton algorithm. This algorithm depends on the first derivative with respect to the parameters. This derivative is **non-linear in the parameters** and highly sensitive to errors in the derivative. Aggravating the difficulty is the fact that, small but opposing errors can produce large errors in the gradient. The smoother the data the smaller the errors in the gradients will be. It is this reduction in error produced by smoothing that allows the Newton optimization algorithm to converge.

LOWESS

The local polynomial smoothing technique used is **LOWESS** (LOcally WEighted Scatter plot Smoothing), see Fan and Gijbels (1996 pp. 24-26). LOWESS is local, in that each data point is the centre of a neighbourhood that has a polynomial regression on the set of points in the neighbourhood. The smoothing is achieved by choosing the polynomials such that points further from the centre have less influence on the fit. Further smoothing is achieved by choosing the polynomials such that they 'smoothly' fit together. Note that the fitted polynomial need not pass through the centre point of a local neighbourhood. The bandwidth f **parameter** of the LOWESS algorithm sets the size of the neighbourhood about a data point for the local polynomial regression.

Smoothing and Bias

The greatest danger of applying smoothing to the original data is that bias may be introduced into the inferred DE parameters. In an **unbiased** procedure the mean of the error must be zero. The error is computed as the un-smoothed data minus the inferred ownership rates. If a bias exits it may be either positive or negative. The smoothing parameters were therefore chosen so as to minimize the square of the mean errors.

Algorithm

The non-linear regression was run for various sequences of f parameters. Figures 1a, 1b, 1c and 1d below plot the mean squared error against the LOWESS $f$parameter. The star point and the doted vertical line indicate the value of minimum squared mean errors, i.e. f*=1.0. Figure 2 shows the observed ownership rate against the LOWESS curve for f* and the projection of the observed flows on to the curve.

Results of non-Linear Regression

The corresponding values of the parameters inferred by non-linear regression are given in table 1 below. The solid curve in figure 7 shows the fit using optimally smoothed non-linear regression. The open circles are the non-smoothed original data points. The inferred values of α and γ are shown as dashed lines. The optimally smoothed non-linear regression curve is a very good fit to the dominant trend of the data.

Table 1

Inferred Parameter Values using f*			correlation
$\hat{\alpha}$	$\hat{\kappa}$	$\hat{\gamma}$	$\hat{\rho}$
437.3824	0.04568628	1979.768	0.9647

The correlation between the original (unsmoothed) and the corresponding values of ownership rate as inferred by non-linear regression is 0.9647. Figures 1a to 1d show the predicted growth in ownership in relation to the original observations.

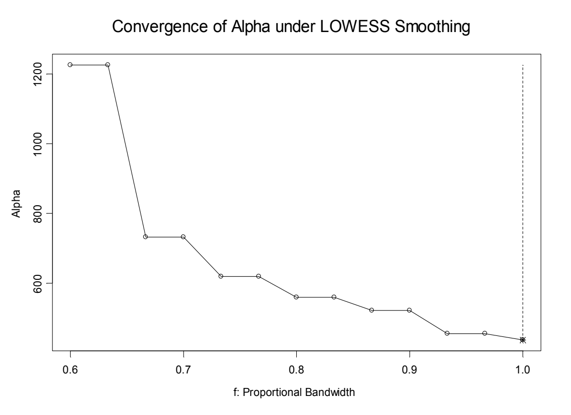

Figure 1a
Minimum bias for α

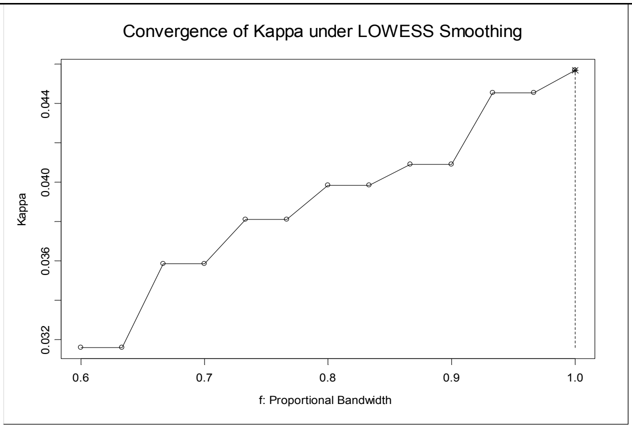

Figure 1b
Minimum bias for κ

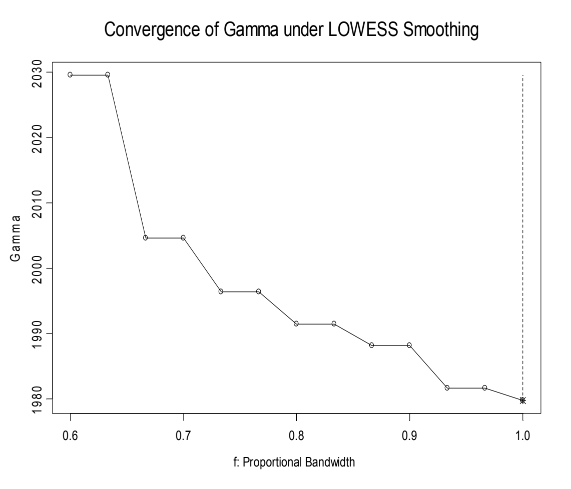

Figure 1c
Minimum bias for γ

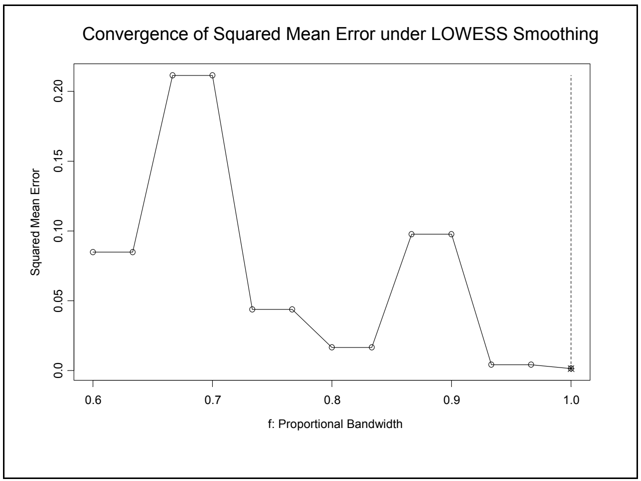

Figure 1d
Minimum Squared Mean Error

ANALYSIS OF RESIDUALS

Residuals

The terminology adopted in the paper is to refer to a residual term rather than an error term. Residuals are a measure of deviation from the 'best' fit of the model to the data. The residual may consists of several components: systematic deviations of the model from reality, or systematic measurement errors, or others components of a random nature. The two properties investigated are bias and trend.

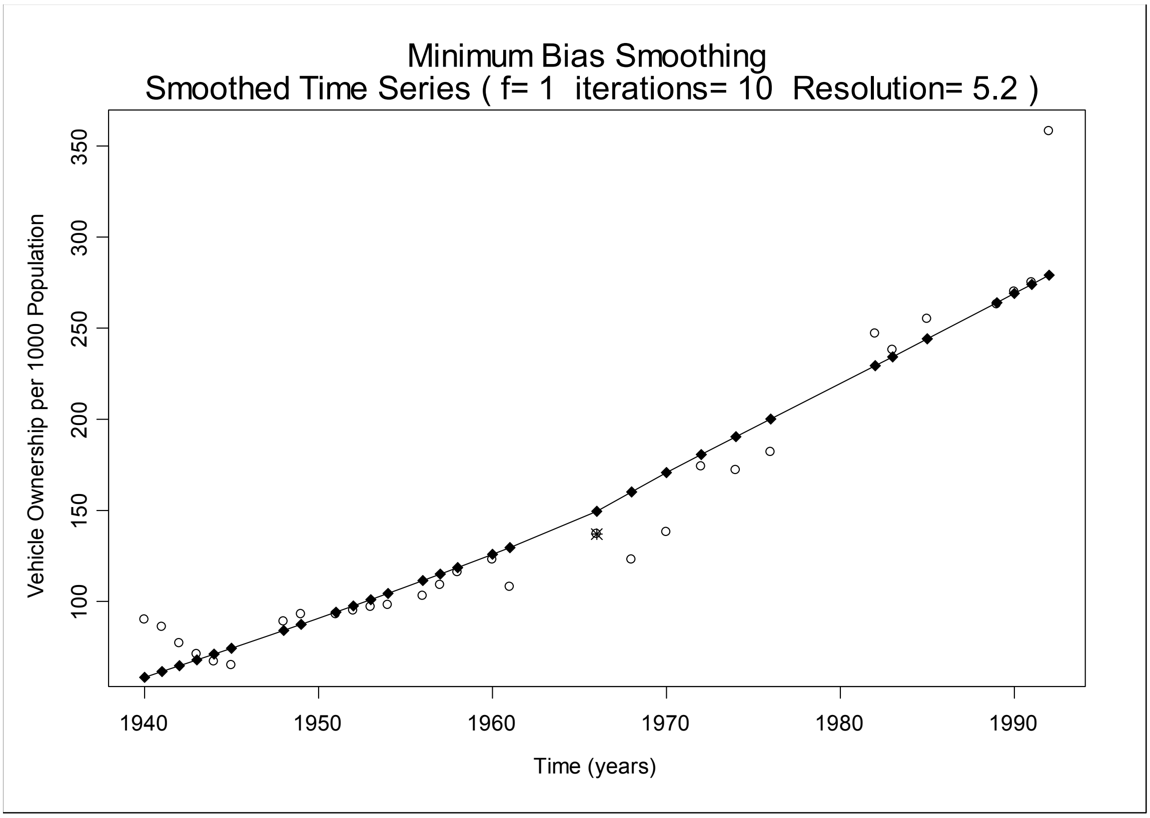

Figure 2 Smoothing Curve giving Minimum Bias

open circles	the observed ownership rate
solid line	the LOWESS curve for f^*
solid diamonds	the projection of the observed ownership rates on to the LOWESS curve

Bias

A major assumption of the non-linear least squares model is that the residual term has a mean of zero. The properties of the residual term are a major indicator of sound statistical inference. Two methods will be used to investigate: the existence of bias in the residual, residual plots and confidence intervals.

Figure 3a is the scatter plot of the residuals. By inspection, the arithmetic mean estimator of bias and the MM robust regression line are indistinguishable from the horizontal axis (i.e. the error equals zero). Hence it may be concluded that there is neither bias nor trend in the error term.

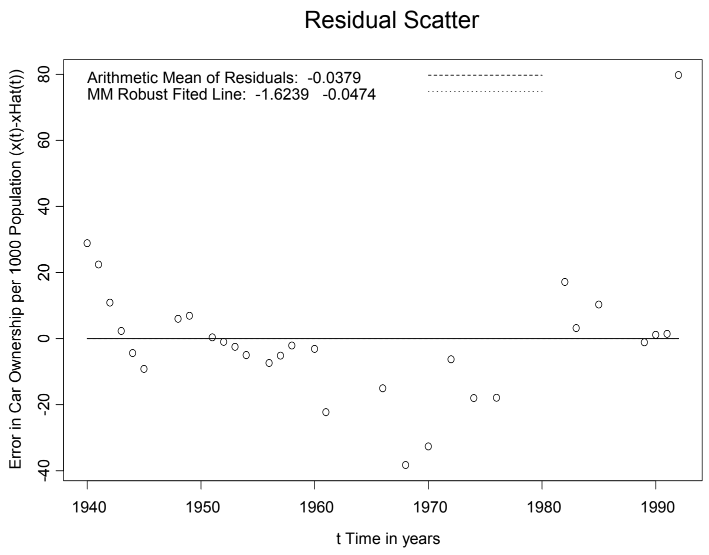

Figure 3a Error Scatter

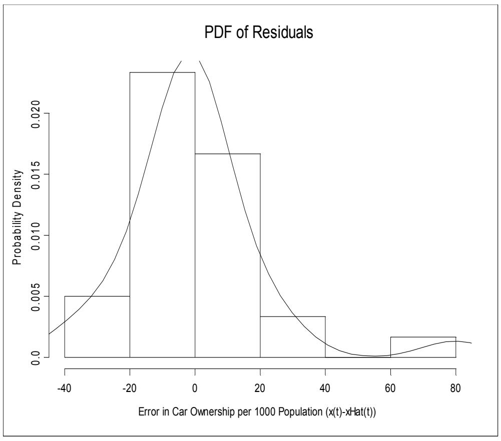

Figure 3b Error Histogram
PDF Probability Density Function

Confidence Intervals

The absolute value of the arithmetic mean, the intercept and the gradient are small compared to the total variation in the data. Despite the small values, further confirmation, beyond the residual plot, may be sought. Such confirmation can be provided by testing whether the bias statistic is significantly different from the global mode (in this case zero). The two-sided test reduces to the problem of constructing a confidence interval around zero. From figure 3b it is apparent that the distribution is both non-unimodal and skewed. The distribution is therefore not a candidate for classical methods of confidence interval construction. Fortunately, bootstrap methods (Efron, and Tibshirani 1993), ("S-Plus 2000 Guide to Statistics vol. 2" 1999) are not as vulnerable to such deviations from the Gaussian distribution. The **bootstrap** methods are re-sampling algorithms in which sapling with replacement from the original sample is used to construct an empirical distribution. Statistics are computed from the empirical distribution from which more robust inferences can be drawn. Both empirical and bias-corrected and accelerated (BCa) algorithms are used.

Empirical Confidence Interval (Percentiles)

The empirical cumulative density function is re-sampled by the bootstrap method, in order to construct a more robust cumulative distribution function (c.d.f.). The inverted bootstrapped empirical c.d.f. is then used to find end points of confidence interval.

BCa Confidence Interval (Percentiles)

The bias-corrected and accelerated (BCa) confidence intervals compensate for possible bias by using the empirical c.d.f.. The acceleration adjustment is used to adjust for deviations from the symmetric Gaussian distribution.

The number of replications re-sampled is one thousand (B = 1000).

The hypotheses to be tested are: $\mathbf{H_0{:}\varepsilon=0}$ **v.s.** $\mathbf{H_1{:}\ \varepsilon\neq 0}$. From tables 2 and 3 below the hypotheses can be tested at the 90% and 95% significance levels.

Table 2

Empirical Bootstrap Percentiles			
2.5% (significance/2)	5% (significance/2)	95% (1- significance/2)	97.5% (1- significance/2)
-6.87	-5.72	6.069	7.311

Since 0∈ (-6.87, 7.311) then **accept H_0:ε=0** at the 95% confidence level.
Since 0∈ (-5.72, 6.069) then **accept H_0:ε=0** at the 90% confidence level.

Table 3

BCa Confidence Limits			
2.5% (significance/2)	5% (significance/2)	95% (1- significance/2)	97.5% (1- significance/2)
-6.495	-5.61	6.36	8.311

Since 0∈ (-6.495, 8.311) then **accept H_0:ε=0** at the 95% confidence level.
Since 0∈ (-5.610, 6.360) then **accept H_0:ε=0** at the 90% confidence level.

It can therefore be concluded that there is no bias in the residual term.

PSEUDO CONFIDENCE INTERVALS

In the previous section confidence intervals for the residual term were investigated. This section treats the problem of constructing confidence intervals for the predicted car ownership. Clearly the usual methods of linear models are inapplicable. Attempting to construct confidence intervals for the predicted points themselves, although formally correct, would involve theoretical contortions best avoided.

Given that the purpose of constructing confidence intervals is to give some indication of the 'accuracy' of the predictions, a compromise problem might be acceptable to the more rigorously defined problem. It is much simpler to construct confidence intervals for the DE parameters, and then plot the upper confidence interval curve as the solution of the DE using the upper extremea of the parameters, and likewise for the lower confidence interval curve.

One shortcoming of this method is that by virtue of the properties of the logistic function, if $\alpha_{upper}>\alpha_{lower}, \kappa_{upper}> \kappa_{lower}, \gamma_{upper}\geq\gamma_{lower}$ then the confidence bound curves intersect. As the pseudo confidence bounds approach the intersection, the size of the true confidence bounds are underestimated. However, as the pseudo-confidence bounds move away from the intersection they approach the true confidence intervals. The pseudo-confidence bound curves form an inner bound to the true confidence interval curves.

Initially bootstrap re-sampling was used to construct a sample of parameter vectors. Unfortunately, the nonlinear regression algorithm did not converge for some samples. The jackknife re-sampling method proved more successful, with the nonlinear regression algorithm converging for every sample. In **jackknife** methods, re-sample from original sample is without replacement. Each re-sample omits a fixed number randomly chosen elements. Note that the non-replacement implies that the re-samples are unique. In order to obtain a large enough number of re-samples so as to construct distribution functions for the parameters, three elements were omitted instead of the usual one element. Five hundred re-samples were generated. This was considered sufficient, since by trial and error it was found that the 5% and 95% confidence end points converged in less than 500 re-samples.

On the each re-sample the parameter is re-estimated and a new cumulative distribution function for the parameter is constructed. The cumulative function can be constructed so that it is strictly increasing and therefore invertible. The 5% and 95% quantiles (end points) can then be determined from the inverse cumulative function. A confidence interval so constructed is known as an **empirical confidence interval**.

In figures 4a, 5a and 6a below, the doted line represents the value of the relevant parameter inferred from the complete sample. The asterisk (*) marks the value of the parameter computed for the initial jackknife sample. CI (confidence interval) Upper Points (solid lines) shows the convergence of the 95% confidence curve as the number of jackknife samples, and therefore inferred values of γ, approaches 500. Likewise, the CI Lower Points (solid lines) show the convergence of the 5% confidence curve. The curve becoming progressively flatter indicates the relevant end point of the confidence interval is stabilizing to a constant.

Figure 7 and Table 4 give the final results of pseudo-confidence interval construction exercise. Figure 7 shows the predicted curve (solid) in relation to the untransformed observations (open circles). The pseudo-confidence interval curves are represented as dots. The inferred values of α and γ are shown as dashed lines.

As would be expected from a reasonable inner bound on the confidence curve, the scatter plot of figure 7 shows that most of the observed points lie within the confidence interval curves. In accordance with the predicted properties of the inner bounds a greater proportion of the points lie within the curves as values of the time variable move away from the cross over point. The exception to this is the World War II data, which contrary to the logistic model decreases due to the rationing of fuel and vehicles during that conflict.

Another exception is the last data point (i.e. point 30, 1992). Its position well out side the inner bound gives further credibility to the belief that it is an outlier. A rapid growth in ownership of 32.3% p.a. seems improbable in economic terms and, by subjective judgement, inconsistent with the logistic model.

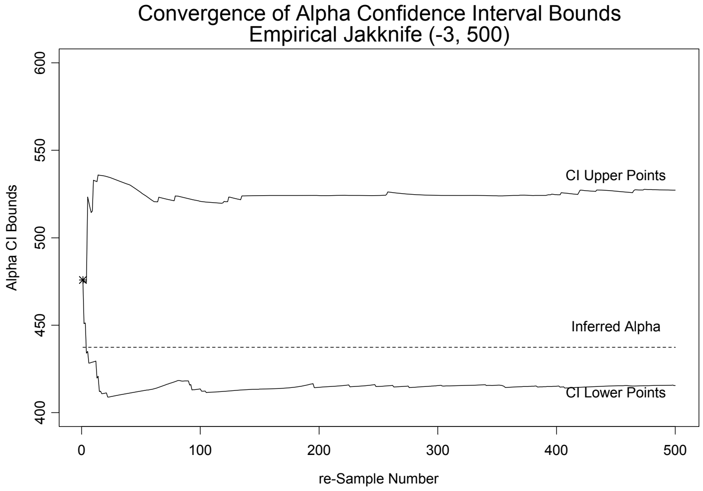

Figure 4a Convergence of Alpha Confidence Interval. The curve stabilizes after re-sample 150.

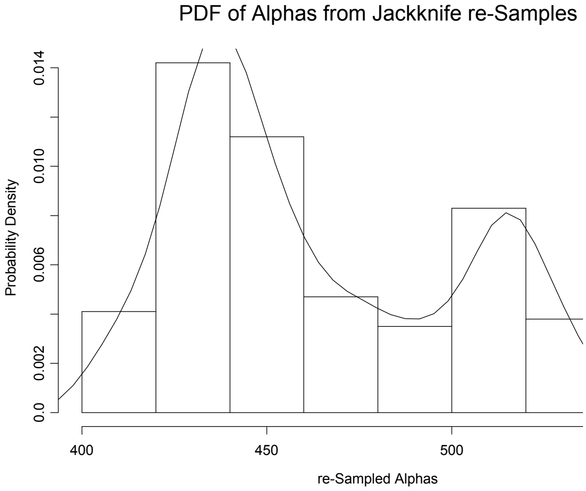

Figure 4b Histogram of Empirical Jackknife re-sample estimates of Alpha

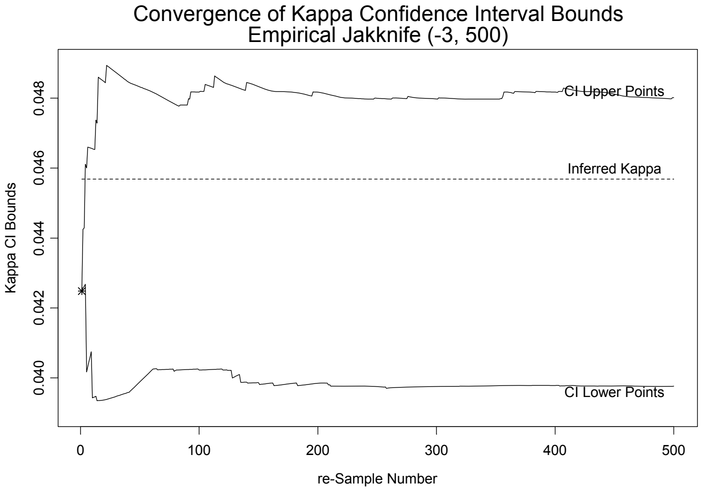

Figure 5a Convergence of Kappa Confidence Interval. The upper curve stabilizes by re-sample 400, the lower curve by re-sample 300.

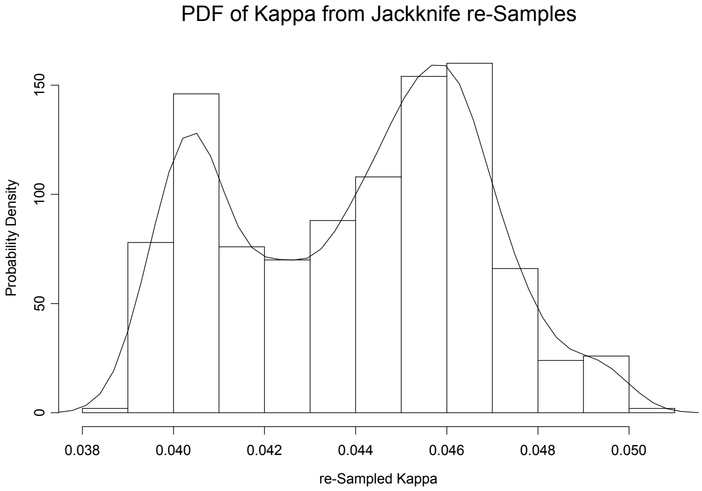

Figure 5b Histogram of Empirical Jackknife re-sample estimates of Kappa

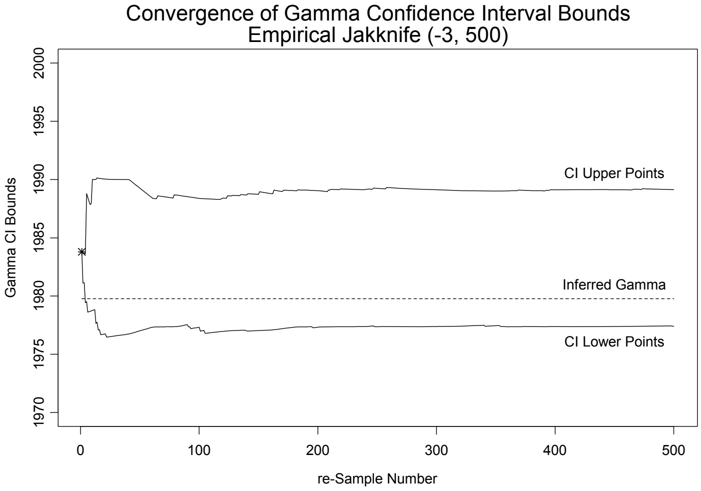

Figure 6a Convergence of Gamma Confidence Interval. Both curves stabilize by re-sample 300.

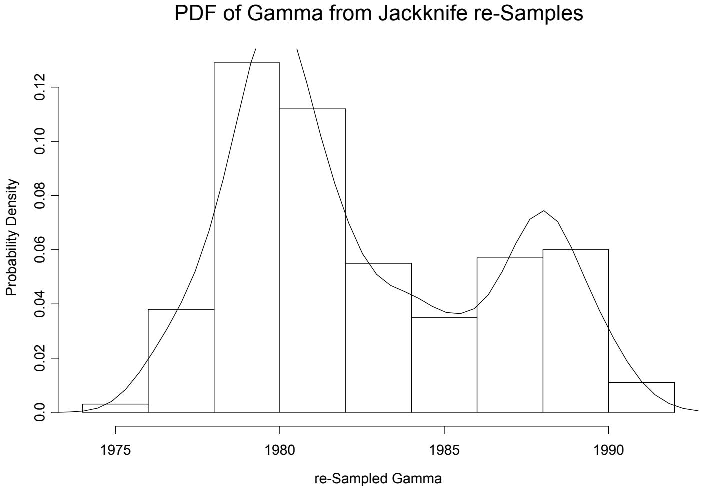

Figure 6b Histogram of Empirical Jackknife re-sample estimates of Gamma

CONCLUSIONS

1. Minimum bias smoothing provides adequate conditioning to allow the convergence of the non-linear regression algorithm.
2. The non-linear regression curve is a good fit to the dominant trend of the data.
3. It should be noted that the boundaries and hence the demographic composition of the CoJ has changed since 1994, therefore this model can only be applied to parts of the CoJ and then only after careful consideration. However, the inference method should be applicable to the expanded CoJ data.
4. The inference process for the non-linear is statistically sound.
5. The fitted curve has high correlation to the raw data.
6. Figure 3a shows that the residuals are unbiased and have a zero trend.
7. The tests of hypotheses based on Tables 3 & 4 demonstrate that the residual is unbiased.
8. The pseudo confidence interval curves give a credible and useful inner bound on the actual 95% confidence curves. These curves converge to stable values using the empirical distribution function constructed by jackknife methods.
9. The algorithm is fully automated in S^+ code.

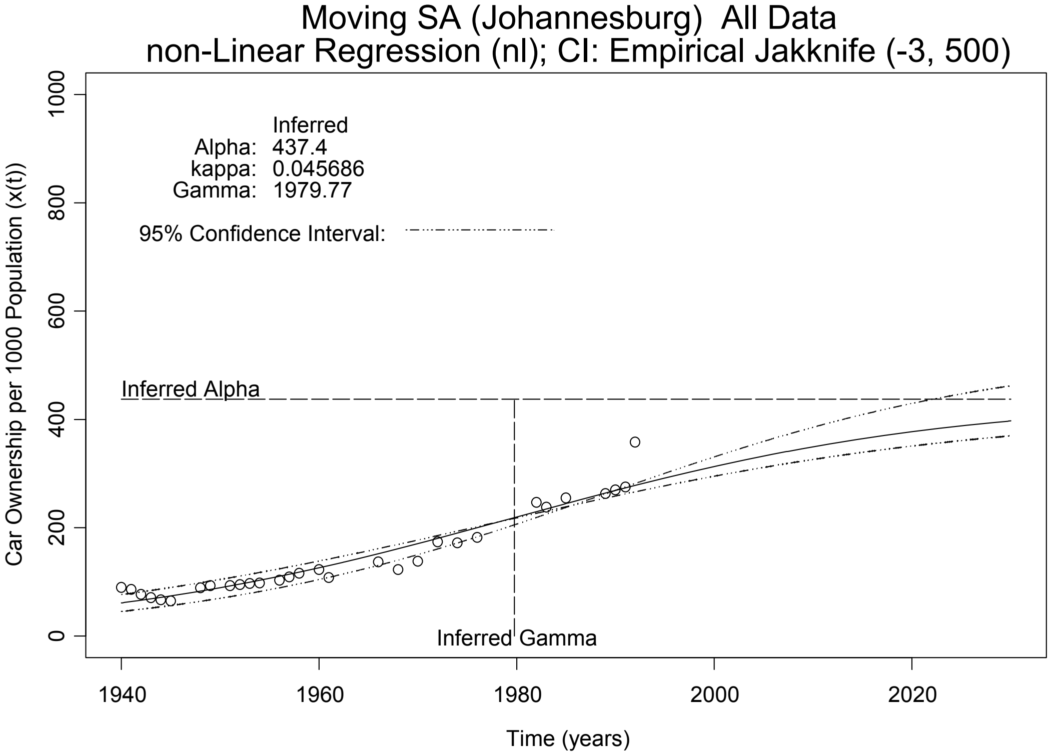

Figure 7: Empirical 95% Confidence Intervals

Table 4 Empirical Confidence Intervals Convergence Value of End Points

Parameters	5%	95%
alpha	415.47	527.18
kappa	0.03976	0.04802
gamma	1977.4	1989.1

ACKNOWLEDGEMENTS

I would like to acknowledge the following individuals and institutions: -

1. The City of Johannesburg for its generous support of this work;
2. e-NaTIS for providing the fundamental component of the research, the data;
3. the libraries of the University of the Witwatersrand and the University of Johannesburg for the privilege of use of their facilities without which this work would be impossible;
4. Professor L. D. Ashwal, of the School of Geosciences in the University of the Witwatersrand, for his invaluable editorial assistance and wise advice.

DISCLAIMER

The views expressed in this paper are those of the author alone and do not represent the opinion of his employer or any other associated institution.

REFERENCES

Efron, B. and Tibshirani R.J. 1993, "An Introduction to the Bootstrap", Chapman & Hall

Fan, J. and Gijbels, I. 1996, "Local Polynomial Modelling and Its Applications", Chapman & Hall

Integrated Transport Plan 2003-2008 vol. 1

Kelly, J., 2007, “Application of Robust Methods to Car Ownership Trends Modelling in Johannesburg”, Proceedings of the Southern African Transport Conference

“Moving SA Car ownership Forecasting” November 1997 Stanway Edwards Ngomane Associates for National Department of Transport (Project 95/093/1)

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

“S-Plus 2000 Guide to Statistics vol.s 1 and 2” 1999, MathSoft Inc.