International Journal of Management, IT & Engineering
Vol. 9 Issue 02, February 2019
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

TO ANALYZE THE INFLUENTIAL OBSERVATIONS IN NONLINEAR REGRESSION

MANISH KUMAR SRIVASTAVA

Manish Kumar Srivastava, Research Scholar, Dept. of Statistic, Himalayan Garhwal University, Uttarakhand

Dr. Pardeep Goel, Professor, Dept. of Statistic, Himalayan Garhwal University, Uttarakhand

ABSTRACT

This work helps to influence nonlinear regression analysis and in particular, to detect influential findings. The emphasis is on a regression model with a defined mean function, which is nonlinear in its parameters and where, according to the knowledge of the data generation method, the function is chosen. In the regression model, the error term is expected to be additive.

The main objective of this study is to carry out diagnostic methods to determine the effect of findings from a nonlinear regression analysis on different outcomes. The results obtained provide diagnostic tools for detecting observations that influence the parameter estimates individually or jointly with some other observations. In addition, it is of interest to determine the conditional effect i.e. the influence of an observation conditional on the removal of another observation. This will help to describe powerful findings that may be overlooked because of complicated relationships between the observations. Novelties of the proposed diagnostic methods include the possibility to assess influence of observations on a given parameter estimate and to assess influence of multiple observations.

A further focus of this study is on the effect of the findings on the result of a testing protocol based on the Rao score test hypothesis. The so-called added parameter plot is a groundbreaking approach to the problem of visual recognition of influential findings concerning the score test statistics obtained in this thesis. New diagnostic indicators are derived as a supplement to the added parameter plot to determine the effect of single and multiple findings on the statistics of the score test.

Keywords: Parameter plot added, approach to differentiation, influential observation, nonlinear regression, score test.

INTRODUCTION

Francis Galton, the English polymath, took a country walk at Na-worth Castle near Carlisle, Northern England, in 1889. The country was swept by a rainstorm, and as Galton took cover from the rainstorm, an idea, namely the idea of correlation analysis, shone across him. According to Barnes (1998), this was the beginning of correlation and regression analysis, where an amusing storey is told about the history of statistics in general and regression analysis in particle form. If this storey is real or not discussed, see for example, Stigler (1986). However, assuming it is real; the notion of correlation that flashed across Galton that day in the rain did not appear in a vacuum. In a 20-year research project, it was a closing move. In the late 1870s, when he performed experiments on seed size in successive gene erations of sweet peas, he first noticed a reversion to the mean. In the 1880s, Galton investigated the heights of parents and their descendants see Bulmer (2003). Galton discovered that as he called it, tall parents, or taller than mediocrity, had

children who were shorter than themselves and that parents who were shorter than mediocrity had children taller than themselves. This led him to name the phenomenon "regression toward mediocrity." The regression phenomenon did not begin with Galton, according to Sen and Srivastava (1990). There were other mathematicians who did what we would call regression before Galton did. With Galton's work, what was important was that he related regression and correlation. Galton was simultaneously conducting two separate studies, one in anatomy and one in forensic science, according to Stigler (1989), in the late 1880s. The question in anthropology was as follows: If a single thigh bone is retrieved and weighed from an ancient grave, what can the bone measurement tell us about the total height of the person to whom it belonged? The other question was related: what can be said about the relationship between measurements taken from different sections of the same person for the purpose of criminal identification? What dawned on Galton was that these new problems were similar to the old kinship problems and that all three of them were a much more general problem, namely that of correlation, in no more than special cases. He not only described the relationships between variables by regression, towards mediocrity, but he also found a way through the correlation coefficient to calculate the intensity of this re-lationship. In addition, Galton noticed that it was possible to break the variance of one variable along the regression line into two parts, one part which could be explained by the other variable and one which could not.

Not far from our conception of them are Galton's ideas about correlation and regression. To study the linear association between two variables, correlation is used. The coefficient of correlation for the calculation of the intensity of this association ranges from -1 to 1, where the sign indicates the direction of the association. The linear relationship between the variables in linear regression is represented by a linear function. In addition, as Galton realised, the dependent variable de- pends on some unnoticeable mistake, often assumed to be a random variable normally distributed with expectation zero and constant variance. A matched linear regression model is obtained when the unknown parameters in the linear function are calculated. As the estimate of the slope parameter in the linear function and the correlation coefficient are functionally related, correlation and regression are associated. Galton correctly predicted that in many applications the regression and correlation methods will have a prominent position, see Bulmer (2003). For example, in business, social and behavioural sciences, biological sciences and many other disciplines, the linear regression model is commonly used.

ANALYSIS IN NON-LINEAR REGRESSION

In this, new methods for conducting influence analysis in nonlinear regression are proposed. The nonlinear regression model referred to in this study is a model where a function which is nonlin-ear in its parameters is the relationship between the variables. We assume that the term of error linearly enters the model. The reason for using the model of nonlinear regression derives from the need for a meaningful and practical model to explain real-life phenomena. The biological, chemical or physical sense may be. Therefore the function used to define the relationship between the variables is known-10 and is chosen because of the knowledge of the data generation process. The model of Michaelis-Menten will be frequently used in this study as an example of a nonlinear regression model. The model is used in the study of enzymatic-catalyzed reactions, called enzyme kinetics, for example. The reason for using it is that the behaviour of the velocity of the enzymatic reaction (dependent variable) is known to be well defined by the Michaelis-Menten model when introducing differ-ent substrate concentrations (independent variable) to the process.

As for linear regression, the current literature on influence analysis in nonlinear regression is not as comprehensive. One explanation for this may be that closed form estimators for the parameters in the nonlinear re-gression model usually do not exist. Cook and Weisberg

and St. Laurent and Cook discuss the identification of influential findings on the fit of the nonlinear regression model. A nonlinear version of Cook's distance was developed by Cook and Weisberg (1982), and St. Laurent and Cook suggested an approach to test the effect of the observations on the fitted values and the es-timate of the variance in a model of nonlinear regression.

INFLUENCE ANALYSIS REGARDING A TEST STATISTIC

Hypothesis testing is an important component of regression analysis. For linear and nonlinear regression, there are several testing procedures available. In this thesis, the emphasis is on the score test.

Several authors have submitted work on the sensitivity of the statistical score test. To test for zero-inflation in count results, Lee et al. (2004) used the score test. The null hypothesis under consideration was that the distribution of the Poisson matches the data observed well. There may however be a large number of zeros in the data for certain applications. In this case, a zero-inflated Poisson model may be a more suitable model. Thus the alternative hypothesis is that a zero-inflated Poisson distribution follows the results. Another score test is also in-terest, namely to test the null hypothesis that the information follows a Poisson zero-inflated model with the alternative that a better model is the zero-inflated negative binomial.

When deleting observations, Lustbader and Moolgavkar (1985) derive an expression for the shift in the statistical score test. For linear regression models, this term is derived, but is addressed in depth in matched case-control studies and survival studies to delete entire risk sets. Matched case-control studies are retrospective, observational studies in which the association between a risk factor and for example, a disease is determined by using a particular matching variable for groups to generate. It is normal to consider adjustments in the score test for the elimination of entire risk sets, i.e. the number of subjects at risk of experiencing a certain incident, with case-control data. In survival analysis, it is more desirable for each person to calculate the diagnosis.

The robustness of score tests for generalised linear re-gression models was discussed by Chen. Robustness against the functional form selected under the alternative hypothesis is the robustness referred to here. Chen also addressed how the score test statistics against potential extreme observations can be made more stable. In addition, the sensitivity of the score test, the Wald test and the probability ratio test in relation to nuisance parameters, That is to say, how changes in the values of the nuisance parameters influence the corresponding test statistics. In addition, for a review of impact research concerning the F-test.

NONLINEAR REGRESSION MODELS AND ESTIMATION

Nonlinear regression models, such as eco-nomics, agriculture and biology, are commonly used in many fields. The decision to use a nonlinear regression model can be taken on the basis of the theoretical knowledge of the data generation method and the likelihood at hand. With the exception of the parameters in the model, the f function is generally completely defined. To the researcher or scientist, the parameters are often significant, where the meaning can be graphical, physical, biological or chemical, for example.

In this thesis we assume that a nonlinear regression model has a known f, and that this function is chosen due to the knowledge about the process generating the data. In its parameters, the regression model is not linear (may be partially linear) and the error term is believed to be additive. The model's general form is

$$y = f(X, \theta) + \varepsilon,$$

where $\boldsymbol{y} : n \times 1$ is a response vector, $\boldsymbol{X} : n \times p$ is the matrix of explanatory vari- ables, $\boldsymbol{\theta} : q \times 1$ is a vector of unknown parameters and $\boldsymbol{\varepsilon} : n \times 1$ is the random error, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$.

THE MICHAELIS-MENTEN MODEL IN ENZYME KINETICS

Consider a researcher who is studying enzyme-catalyzed reactions. The scientist knows that an enzymatic reaction's initial velocity fits Michaelis-Menten kinetics. That is, the Michaelis-Menten equation models the relationship between y, the initial velocity of the enzymatic reaction, and x, substrate concentration.

$$f(x) = \frac{V_{max} x}{K_{max} + x},$$

Where Vmax and Kmax are unknown parameters, which will later be clarified In this case, the function f is known to the scientist due to knowledge of chemical reactions, and there is no need to look for the correct functional relationship between y and x. See Briggs and Haldane for more information about the theoretical basis of the Michaelis-Menten equation (1925). By assuming an additive error term, the Michaelis-Menten equation can be used to formulate a non-linear regression model.

$$y_i = \frac{\theta_1 x_i}{\theta_2 + x_i} + \varepsilon_i, \quad i = 1, \ldots, n,$$

Where1=Vmaxand2=Kmaxandεii.i.d. N(0,σ2). Interested readers may see Richie and Prvan, Pasaribu and Dette and Kunert (2014) using the Michaelis-Menten equation forstatistical analysis of enzyme kinetics results. It can be seen that the parameter enters the model linearly, but the parameter enters nonlinearly and hence the nonlinear relationship between yandxis. There are physical representations of the parameters 1 and 2 in the model (2.4). The parameter is the overall initial velocity that is theoretically obtained with respect to the concentration of a substrate when the enzyme is saturated. The second parameter,' 2, is the Michaelis parameter, which is equal to the substrate concentration for the initial velocity of "half-maximum" They are based when the parameters are calculated in the model, since a change in 1 results in a change in 2 as well.

THE GOMPERTZ GROWTH CURVE MODEL

In microbiology, models are also used under various physical or chemical conditions to explain the behaviour or development of microorganisms. Growth is calculated and modelled in order to create these models. It is popular for this reason to use a type of nonlinear models called models of the growth curve. There is a particular interest in the Gompertz growth curve model, and it is defined as

$$f(t) = k \exp(-\exp(a - bt)),$$

Wheref (t) is the size of the timetable population, andk, and unknown parameters. For a discussion of modelling bacterial growth with growth curve models, see e.g. Zweiteringet al. (1990) and Chakrabortyet al. (2014) forstatistical analysis of the Gompertz growth curve model. A common feature of the models of the Gompertz growth curve is that they have two asymptotes: the curve reaches zero asx, a positive constant asx, and accelerates to a maximum value. The point (x,f) where the growth rate is maximum is referred to as point The curve approacheskasx, for the Gompertz growth curve model another characteristic of the Gompertz growth curve model is that it is not symmetrical at the inflection point. Bacterial growth also shows a stage in which the real growth rate begins with a zero ata value. In a given period of time, the growth rate then accelerates to a maximum value,

resulting in a so-called lag time. The growth curve then reaches a final stage in which the growth rate decreases and gradually reaches zero. The size of the bacteria population then approaches an asymptote. Different models of the growth curve can be used to explain bacterial growth activity. However, parameters that are micro-biologically important should be included in the most acceptable model. One of the candidate models used is the modified form of the Gompertz growth curve, which is given by

$$f(t) = A \exp\,[-\exp\,[(\mu_m e/A)\,(\lambda - t)\,+1]],$$

Where the asymptote asx is the lag time is, and in a certain period of time the overall growth rate is μmis. By considering a tangent line at the point of inflection, an alternative definition of λ and μ given the μm parameter is defined as the tangent line slope and the λ parameter is the tangent line intercept.

Graphical displays

Graphical displays are commonly used in linear regression as diagnostic tools and have a long history. We can go back nearly a decade and find strategies that are still in use today: the partial residual plot suggested by Ezekiel, for example (1924). Nevertheless a considerable amount of work has been undertaken to develop current graphical diagnostic tools and to build new ones. A few accomplishments worth noting in the field of regression graphics are the evolution of the use of ordinary residuals in different scatter plots in the 1970s towards the use of different standardised residuals, see e.g. Behnken and Draper (1972) and Andrewsand Pregibon (1972) (1978). The added variable map, a diagnostic instrument that can be used in multiple linear regressions, was suggested by Mosteller and Tukey (1977). The next section will explain this plot more thoroughly. See Cook (1998) for more references on regression graphics. In the case of linear regression, research on graphical instruments in nonlinear regression has not been as comprehensive. Cook (1987) brought new concepts and creative ideas into the field, where the plot was suggested for use in nonlinear regression, similar to the added variable plot.

GRAPHICAL DISPLAYS IN NONLINEAR REGRESSION

It is possible to expand the ideas behind the AVP mentioned to non-linear regression models. A plot similar to the AVP defined by Cook is known as a first-order extension of an AVP. In the next section, a further discussion about this plot will be given, where we also derive one of the key results of this study, the added parameter plot, APP, along with its features that distinguish the APP from the first-order extension of an AVP. The following re-mark is necessary before embarking on the thorough discussions of the APP. Note that the AVP is intended to show available information to determine the importance of a particular explanatory variable for the linear regression model. In addition, the AVP is used to identify observations that have a significant effect on the corresponding total added variable parameter calculation. Because of the one-to-one correspondence between the variables and the parameters in the model, the relationship between the added variable Xp and the parameter estimate is possible. It should however be remembered that this one-to-one correspondence does not usually occur in models of nonlinear regressions. Therefore the fundamental goal of the nonlinear regression plots comparable to the AVP is to show information related to the inference of the selected parameter rather than the selected variable.

EVALUATING MODEL ASSUMPTIONS

The next step is to test main model assumptions when we deal with one model: normally distributed errors, independent errors and homogeneous variance for errors. This phase and the steps below are not special to nonlinear models, but are common to all linear models. Substantial deviations from the assumptions could lead to bias (inaccurate estimates), standard errors that are skewed, or both. From an analysis of the residuals, violations of these assumptions can be identified by graphical procedures and systematic statistical tests.

For a detailed analysis.Briefly, to check if the distribution of the measure-ment errors meets normality, the standardised residual plot is commonly applied. The main causes of deviations from normality are outliers and many extreme values. By looking at the plot of the fitted values over the residuals (absolute residuals, which are raw residuals stripped of the negative sign, or standardised residuals, which are raw residuals scaled by the variance; see the illustration below), heterogeneity of variance can be observed. When the residual errors display a pattern (e.g., increasing variability as the explanatory variable rises. In our example, this is the case for. The parameter estimates may not be significantly affected if variance variability is overlooked, but this may lead to severely misleading confidence and prediction intervals. The residuals are believed to be independent, and it is visually obvious in a plot of correlations of residuals against "lag" when this assumption is violated (or units of separation in time or space). Variables calculated over time on the same subject (e.g., plant, animal, or soil sample) usually appear to result in autocorrelated residuals that need to be compensated for by modelling the matrix of variance-covariance.

WHY SHOULD WE USE NONLINEAR MODELS

Parsimony, interpretability, and prediction are the main benefits of nonlinear models. In general nonlinear models are capable of accommodating a wide range of mean functions, but in terms of the variety of data they can describe, each individual nonlinear model can be less versatile than linear models (i.e. polynomials); however, nonlinear models suitable for a given application can be more parsimonious (i.e. there would be less parameters involved) and more readily integrated. Interpretability stems from the fact that a biologically significant mechanism can be correlated with the parameters. For instance, the logistic equation is one of the most commonly used nonlinear models (Eq. [2.1] in Table 1). The prevalent S-shaped curve of growth is represented in this model. The parameters have a simple sense (see Table 1) and their description is correlated with units. The asymptotic parameter (Yasym) has units that are equal to the response variable (Y), the inflection point (tm) has units that are equal to the independent variable (t), and the curve steepness arameter (k) has units that are equal to t. This last parameter can be interpreted as the time it takes to travel from the inflection point to around 0.73 of the asymptotic value (when t is the time). A competing polynomial model used to explain the same data would have the drawbacks of having more parameters (more than just three) and of not easily understanding the parameters. For instance, what would be a five-degree polynomial interpretation of the parameters? The ultimate benefit of using nonlinear regression models is that their forecasts appear to be more robust than competing polynomials, especially beyond the range of data observed (i.e., extrapolation). However, nonlinear regression models come at an expense. Their key drawbacks are that they can be less versatile than competing linear models and that no analytical approach for parameter estimation is typically available. As a result, the first argument is that model choice is important. It is tempting to then try a large library of functions and select the model with the lowest error; however, it is almost always easier to choose a model based on whether it has been used effectively in similar applications and has biologically mean-ingful parameters.The lack of an empirical solution has two practical conse-quences. First to find estimates for the parameters, a numerical approach must be used, and this means that the algorithm's convergence needs to be verified. The second factor frequently arises from a lack of convergence, which is that these numerical approaches need starting values. Choosing a model with biologically relevant parameters makes it easier to choose starting values since the starting values can typically be easily calculated by visual data inspection.

CONCLUSION

It is well known that not all findings play an equal part in evaluating the different outcomes of a regression study. For example, only a few observations can decide the character of the

regression line, while the data is somewhat ignored. These findings that significantly impact the results of the study are considered influential observations. There is a large collection of diagnostic methods to use for finding influential findings following the linear regression model. For nonlinear regression models, the amount of literature and research effect analysis is not as comprehensive as in the case of linear regression. We want to contribute to the effect study of different outcomes of the nonlinear regresion analysis with this dissertation. In particular, when testing a hypothesis that a specific parameter in the nonlinear regression model equals zero, we focus on the task of finding findings with substantial impact on the parameter estimates of a nonlinear regression model and on the score test statistics. The main contributions of this study are as follows: Two separate diagnostic tests to determine the influence When we are interested in evaluating the effect of an observation on the entire vector of parameter estimates, the first calculation, DIM, k, should be used.

As compared to joint influence, multiple observations may exercise whatwe refer to as conditional influence on the parameter estimates. If an observation is not marked as influential, conditional effect exists unless another observation is deleted first. Therefore an impact measure is suggested to determine the effect of thekth observation, provided that theith ob-servation is deleted. To calculate the effect of the findings on the statistics of the score test, we suggest two measures of influence. When evaluating the effect of a single observation on the statistical score test, denoted DIMSk, the first calculation is to be used. In general, we are proud to propose our new measures and diagnostic methods, as they add to the study of nonlinear regression influence analysis. With this thesis, we give practitioners more approaches to choosing from whencoco. However we want to highlight some of the contributions we feel especially good about first, the use of the proposed marginal impact tests, DIM j, kandDIM j, K, offers the ability to test the impact of findings on a specific estimate of parameters. Diagnostic measures exist to determine the effect of findings on a particular parameter estimate in the linear re-gression model, but parameter estimates in the online regression model have not yet been performed. Secondly, the nonlinear1466 parameters estimate Regression models are complex because the estimators usually do not have a closed form. Instead the estimates must be found using iterative methods. Take a contrast and recommend following the case-deletion approach to test the effect of findings on parameter estimates (or other statistics which are functions of the parameter estimates). With this strategy, we need to iteratively find the estimates for each deleted observation. This might become a daunting task. Our suggested approach to analytical impact decreases the burden of additional iterations, because we need to find the parameter estimates only once. In addition, after reviewing the literature on the study of impact in nonlinear regression, we have not yet seen any research findings on how observations that are influential on the outcome of a hypothesis test procedure can be established. With the proposed findings, the added parameter plot and the DIMSk and DIMSK diagnostic tests, we offer another view of nonlinear regression effect analysis, as most of the study focuses on estimates of parameters.

The most important step that separates nonlinear models from linear models is that without sufficient guidance, the choice of the main function is crucial and this can be difficult. We have provided a comprehensive library of nonlinear functions and typical applications that, we hope, will make it easier to choose candidate models. Our analysis of nonlinear equations is incomplete since there are countless possible function numbers to be used and ad hoc modifications. In order to avoid common errors in the use of nonlinear regression models, we have also provided a suggested work flow that should provide the required structure.

REFERENCES

- Alfons, A., Croux, C. and Gelper, S. (2018). Inadequate least managed squaresregression for investigating high-dimensional enormous information sets.The Annals of AppliedStatistics,7, 226-248.
- Andrews, D.F. and Pregibon, D. (2018). Finding the anomalies that matter.Journalof the Royal Statistical Society. Arrangement B,40, 85-93.
- Atkins, G.L. and Nimmo, I.A. (2017). An examination of seven techniques for fittingthe Michaelis-Menten equation.Biochemical Journal,149, 775-777.
- Atkinson, A.C. (2017). Relapse diagnostics, changes and constructedvariables.Journal of the Royal Statistical Society. Arrangement B,44, 1-36.
- Atkinson, A.C. (2016).Plots, Transformations and Regression. Clarendon,Oxford.
- Atkinson, A.C. (2016). Concealing unmasked.Biometrika,73, 533-541.
- Barnes, T.J. (2014). The historical backdrop of relapse: entertainers, organizations, machines andnumbers. Environment and Planning A,30, 203-223
- Bates, D.M. and Watts, D.G. (2014).Nonlinear Regression Analysis and ItsApplications. Wiley, New Jersey.
- Behnken, D.W. and Draper, N.R. (1972). Residuals and their fluctuation patterns.Technometrics,14, 101-111.
- Belsley, D.A., Kuh, E. and Welsch, R.E. (1980).Regression Diagnostics: Identi-fying Influencial Data and Sources of Collinearity. Wiley, New Jersey.
- Beyaztas, U. and Alin, A. (2014). Adequate folding blade after-bootstrap methodfor location of persuasive perceptions in direct relapse models.StatisticalPapers,55, 1001-1018.
- Briggs, G.E. and Haldane, J.B.S. (1925). A note on the energy of catalyst action.Biochemical Journal,19, 338-339.
- Bulmer, M. (2003).Francis Galton: Pioneer of Heredity and Biometry. JohnHopkins University Press, Baltimore.
- Chakraborty, B., Bhattacharya, S., Basu, A., Bandyopadhyay, S. and Bhattachar-jee, A. (2014). Decency of-fit testing for the Gompertz development bend model.Metron,72, 45-64.
- Chatterjee, S. and Hadi, A.S. (1986). Compelling perceptions, high leveragepoints, and exceptions in straight regression.Statistical Science,1, 379-393.