Nonlinear Estimation from Scratch

John C Nash

2025-04-13

Motivations

In early 2025, Nasir Bashir of Cambridge University emailed me with a question about measures of bias along with variance in estimated nonlinear parameters. Frankly, I have never paid a lot of attention to such matters. The package *nlsr* of which I am the maintainer and a major developer uses the standard linearization approach to computing estimated standard errors of the parameters. These are, of course, a heroic (pseudo?) approximation, though what they approximate is open to debate.

As I started to read some of the papers on measures of bias for nonlinear parameters, for example, Box (1971), Philip Hougaard (1982), P. Hougaard (1985), I was discouraged by the need to make a number of assumptions I consider to be tenuous and the truly Byzantine notation used by most authors.

In a pleasant (jazz) concert given in a local church on a Sunday afternoon, my thoughts coalesced into a moderately organized blob. This article is an attempt to delineate those thoughts, hopefully in a way that is useful to me and to others.

In advance, I will apologize for the notation used. I have aimed for simplicity in symbols and, where possible, avoided Greek symbols, special letter markings, super- and sub-scripts and other notations that are awkward to enter and print. This is certain to offend at least one group of workers. I nevertheless hope my exposition is readable.

What is model estimation?

Modelling or **model estimation** is a prominent activity of statisticians and others who use statistical tools and methods. Unfortunately, we – and I include myself – frequently apply "standard" methods, often with much less attention to why we do so.

All modelling starts with some data and some metadata. Let us denote our collection of data as M. The metadata will, for the moment, be labelled as `story`. It may appear in various guises, many of which are non-mathematical. I tend to think of M as a table of numbers and text with rows being observations and columns being quantities observed. Sometimes we call these **variables**, but some of the columns may be tags that are, in fact, part of the metadata. One of the most common situations has a column of M, call it m, as the quantity we wish to predict, or model, based on one or several of the other columns.

We ideally (including here) presume that there is a set of true data T and some true but unknown function F_{true} of some true but unknown parameters β such that

$$m_{true} = F_{true}(\beta, T)$$

More generally, we might seek a model that tries to simultaneously predict all the **variables** in M if we knew T.

Sources of data

We will assume that data is observed or measured by some process. Hopefully we can document the process so that we can eventually uncover deficiencies in the process or mistakes in its application.

What we don't know (can hurt us)

We don't know T. Our first bit of handwaving is to presume that

$$M \approx T$$

However, there are many, many sins that live inside that $\approx$. Let us try to give at least a bit more attention to this.

First, let us look at just one column of data, m, and let $t = m_{true}$.

The **error** in m, our **measured data**, can be written

$$e = m - t$$

If we presume some process $P()$ that carries out measurements, then

$$m = P(t)$$

Statisticians like to make the very helpful assumption that $P()$ is such that the errors are **independently and identically distributed**. They (including I) almost always assume that the center, generally the mean, of the distribution is 0. We write

$$e \sim D(parameters)$$

Very commonly D is the Gaussian distribution with 0 mean and standard deviation σ, thus

$$e \sim N(\mu = 0, \sigma^2)$$

All this is very nice, but ...

The error distribution may not be Gaussian

In most real-world situations, especially those involving human behaviour, there is no reason to think errors conveniently distribute themselves according to a Gaussian distribution.

The errors may not be independently distributed

In many cases we may expect errors to have properties that are linked to one or more of the columns of M, or to something we have not observed at all. They may also be linked to each other in some way. For example, a big error at one time point may imply a big error in the next few. This is serial correlation, but that is but one possible pattern.

Errors may be the result of perfidy

The people recording the data may not be honest. Or they may be misguided or misinformed so that the measurements are improperly recorded. I thought the word 'perfidy' the right one for such errors. In particular, I recall Jane Gentleman presenting an informal talk, I believe to the Statisical Society of Ottawa, entitled "Data's Perilous Journey" where she told how one important Statistics Canada series of death and morbidity data was corrupted by an over-zealous clerk who tagged observations incorrectly out of misplaced moral fervour against drinking. The mis-recording was noted **after** the raw data sheets had been destroyed.

I recall the tagging attributed deaths and injuries to driving while under the influence of alcohol when there was not actual documentation of this.

Errors may be the result of sloppy work or programming

In my first "real" job after getting my doctorate, I was engaged as an Economist Statistician with Agriculture Canada. Soon after I arrived, my keyboarding skills were dragooned one Sunday afternoon to assist with a panic effort to get results from an econometric model. This was the age of Fortran and punched cards, and I was salaried, whereas keypunch clerks were wage employees with a "triple time on Sunday" clause.

After entering 20 numbers of a time series, I refused to continue without an explanation of the data source, since I could demonstrate that there was only a 1 in 10^8 chance the data was not corrupted. My reasoning was as follows:

- The final digits of all entries ended in 3, 5, 8, or 0.
- The change of this happening at random are 0.4^{20} for 20 entries.
- I knew that $log_{10}(4) \approx 0.6021$ (This knowledge was surely an artifact of being educated before electronic calculators were available. At the time I knew a fair number of entries in log tables.)
- $20 * 0.6 + 20 * (-1) = -8$ is a good approximation to $log(0.4^{20}) \approx -7.9588$

It turned out the data were data converted from annual to quarterly figures and "rounded" (always upwards) to 1 decimal. This explains but does not necessarily excuse the use of such data, or the preparation of a quarterly time series from annual data in the way that was used.

Data Mistakes

Data also suffers from what I term "mistakes". Fumble-fingers on keyboards and dyslexia cause the wrong or transposed digits to be recorded. Such errors may be random, or may be correlated with how, when and by whom the measurements are recorded.

Such errors may be large. Clearly they are almost certain to be outside an anticipated distribution of errors and can distort the estimated model, but in the estimation process the presence of the mistake may be camouflaged.

Metadata and its origins

The `story` should tell us about both the observations (rows) and quantities / variables. Ideally, this should tell where, when, how and by whom each observation has been observed or measured, but that is very rare. Each column quantity should also be described and explained.

Beyond describing M, metadata should tell us whatever is known about the model appropriate to the phenomenon under study, including limits and other constraints. This may include bounds or other relationships imposed on parameters of hypothesized models, though there is not a guarantee that the model form is correct.

In full generality, the `story` can include various attempts to model the data, to identify particular mistakes or errors, and understand the processes that produce M.

Components of error

- random distributional errors in the measurement process
- systematic errors in the measurement process, either by poor design or "perfidy"
- mistakes

How estimation and modelling are carried out (nonlinear or other) should be conditioned on our knowledge or assumptions of the presence and relative size of the different errors. Many statistical methods only acknowledge the first category of error, and furthermore may assume a particular distributional form, or at least some distributional properties like symmetry.

Estimation methods for models

In order to model data we need

- a model form, generally in the form of mathematical functions, though possibly expressed in the form of an algorithm or computer program
- some parameters that modify the particular value of the computed model

There are a number of methods used to estimate models, of which the two major themes are

- numerical optimization of an objective function, possibly subject to constraints, by modifying the parameters. There are many algorithms, as well as ways of starting the iterations that such methods employ.
- averaging or filtering methods, generally derived from applying the same principles as used in the optimization above by applying calculus and algebra analytically. In these methods I include neural networks and similar filters.

Issues in parameter estimation

The issue that I believe is the most troublesome in statistical modelling is that of the choice of the model form. This may not be critical is we want only to provide a useful short-hand summary of the data, that is, if we only want some low-storage way to generate the observed data quickly to within some tolerance. This is akin to the interpolation tools used to compute values for transcendental functions by use of polynomial or rational functions. In such cases, most of the additional statistical machinery concerning the possible error in the estimates is not usually relevant.

On the other hand, extrapolating a linear model of a process that is exponential will soon get us into big trouble. The model may possibly be useful for some applications over the range and domain of the data used in the estimation. It may be dangerous to apply outside of this domain.

Assuming we have a model form that is appropriate – and there may be more than one form for a given situation – we then must deal with some other matters.

Multiple solutions Optimization approaches to estimation reveal the possibility of multiple (local) minima. Workers trained in linear modelling, especially if they are urged to not use models where the design matrix is singular, are frequently blind to this possibility. Yet even for linear modelling it is possible to compute models that fit the data but are not unique. Tools (e.g., the singular value decomposition) have made it possible to compute such models safely for over half a century. The difficulty is not in the fit but in the understanding of the meaning of the parameter values.

In nonlinear modelling, understanding the parameter meaning is already demanding. Here, it is often less obvious that there are multiple minima, and they may not have equivalent fit (that is, the loss function or deviation between model and data may not be the same for the different local minima).

Singularities in the models Nonlinear models may not always be defined in regions of the parameter space where estimation procedures try to explore. For example, the logarithm function is only defined for positive values.

Constraints Statistical methods have traditionally avoided the application of constraints. It is my view that the general approach is to carry out modelling or estimation and then see if the constraints are satisfied – a "keep fingers crossed" approach.

Algorithms and Stabilizations There are many algorithmic approaches to estimating models. Here we will only consider nonlinear model tools. Most workers rely on the tools built into the software they use to carry out statistical calculations. These may, as in R, be intertwined with various other tools so that it is difficult or impossible to substitute other algorithms. Thus `nls()` in R uses a fairly simple Gauss-Newton algorithm. We have pointed out several deficiencies in this approach (Nash and Bhattacharjee (2024)), but

a rich set of accompanying tools is available. So far we have not found a satisfactory way to substitute a stabilized (that is, using ideas of Marquardt (1963)) algorithms to avoid some of the unnecessary "singular gradient" error messages.

Our own `nlsr` package (John C Nash and Duncan Murdoch (2019)) uses stabilizations and, moreover, tries to employ symbolic or automatic differentiation to get the necessary Jacobian matrix.

A further algorithmic issue is that tools may be designed to compute a large and flexible infrastructure, then call parts of this as needed. This is the `nls()` approach. By contrast, `nlsr` employs a "just in time" philosophy. In this situation, we have found these two viewpoints do not mesh well.

Programming choices (tolerances and similar needs) Nonlinear models are generally more difficult for human thinking to visualize because the scale changes over the domain of the parameters. Linear models have the output (i.e., the dependent or modelled variable) change in fixed proportion to changes in each input (or independent variable). This is definitely not the case for many or perhaps most nonlinear models.

One consequence of this is that measures of no change or "convergence" can be very difficult to set outside the context of a particular problem. Over many years I have found it important to try to avoid providing fixed tolerances. Unfortunately, floating point numbers can be awkward to compare, particularly if they are very small or very big. An alternative which I prefer, though it is by no means a panacea, is to compare two numbers after adding some "offset" value such as 100. This is still far from perfect, but does avoid some issues, though it may cause programs to keep trying to improve parameter values to seek minima or maxima much longer than is truly necessary. It is worth remembering that

```
Algorithms converge but programs terminate
```

In some cases it is worth customizing the tests for particular situations, but that requires access to the code and ways to make the changes easily.

Estimation of variation in or about a model

The general approach to estimating the "error" in a model is to use the residual sum of squares. However, it is worth remembering that we usually have minimized this quantity with respect to the model parameters. This means the minimization process will have "chased" large individual errors in data, if they are present, thereby distorting the model. Nevertheless, we will still used the residual sum of squares as the basis for estimates of error around the proposed model, that is, form and particular parameters.

We note that our estimates of the variance or standard deviation around a model may be biased. This is akin to the average squared deviation from the sample mean giving a biased estimate of the variance, which is corrected by division by `(n -1)`, where `n` is the sample size. (https://en.wikipedia.org/wiki/Bias_of_an_estimator)

Estimation of variation in model parameters

It is traditional to try to provide "standard error" values for parameters, and hence suggest t-values for the parameters to judge their reliability. Truthfully, for nonlinear models, it is misguided to trust such quantities too far.

These t-values depend on

- the model form being correct
- systematic or mistake errors being "small" and therefore capbable of being ignored
- being able to solve / transform the measures of variation around the model into distributions around the proposed parameter values.

This last issue relates to using the Jacobian in place of the linear modelling design matrix X and being able to assume that we can make a linear approximation "near" to the proposed set of parameters. (?? a reference here would be useful)

Carrying the process further, we may want to consider measures of quality for the variance of the parameter estimates. This is where this article started, i.e., with Box (1971), Philip Hougaard (1982), and P. Hougaard (1985).

?? How to carry this forward.

Diagnostic concerns

A different direction of investigation concerns how we might detect systematic errors or mistakes.

How do we test for the wrong model

Given that getting the wrong model form could be a disaster, especially if we want to extrapolate outside the data domain, consideration of tests for the wrong form would be useful. This exercise will be made more difficult by the power of optimization techniques to "pull" the false model towards the data.

?? ideas

Testing for failure of the iid assumption

Assuming there are few or no systematic errors or mistakes in the data, we can consider trying to test for failures in assumptions, in particular the independent and identically distributed (iid) error assumption.

The two common classes of failures in the iid assumption are

- Autocorrelation: that is relationships between error (or even residual) values that depend on the value of one of the independent variables or on the observation number.
- Heteroskedasticity: that is, unequal variation of errors similarly related to the data.

If we believe that the model form is correct, then the residuals should allow us to use known approaches to explore these violations of assumptions.

?? Durbin-Watson, ?? others

Testing for shape of error distribution

The traditional measures and tests of parameter variation (standard errors and t-statistics) are based on linearization of the model with an assumption of iid Gaussian errors around the model. The linearization and iid assumptions are important, but for nonlinear models involving functions that are bounded in some way, such as logarithms, the Gaussian assumption may also fail. If we have sufficient data and sufficient faith that the other assumptions are met, we could test the distribution of the residuals.

Testing for mistakes or systematic measurement errors

If we believe that there may be mistakes or systematic errors in the data, we need to propose some ways to detect them. As far as I am aware, such detection schemes are almost always customized to the situation at hand, for example, the "last digit" approach above, even if the underlying principle is to look for unlikely patterns in the data or residuals.

We may be able to take advantage of resampling methods (https://en.wikipedia.org/wiki/Resampling_(statistics)). Generally, these methods assume the data is "correct", but outliers in estimated models may suggest errors in some data values, though they are only a beginning of a full diagnosis.

?? how to extend

Discussion

?? Summarize key ideas

References

Box, M. J. 1971. "Bias in Nonlinear Estimation." *Journal of the Royal Statistical Society: Series B (Methodological)* 33 (2): 171–90. https://doi.org/10.1111/j.2517-6161.1971.tb00871.x.

Hougaard, P. 1985. "The Appropriateness of the Asymptotic Distribution in a Nonlinear Regression Model in Relation to Curvature." *Journal of the Royal Statistical Society: Series B (Methodological)* 47 (1): 103–14. https://doi.org/https://doi.org/10.1111/j.2517-6161.1985.tb01336.x.

Hougaard, Philip. 1982. "Parametrizations of Non-Linear Models." *Journal of the Royal Statistical Society: Series B (Methodological)* 44 (2): 244–52. https://doi.org/https://doi.org/10.1111/j.2517-6161.1982.tb01205.x.

John C Nash, and Duncan Murdoch. 2019. *nlsr: Functions for Nonlinear Least Squares Solutions*.

Marquardt, Donald W. 1963. "An Algorithm for Least-Squares Estimation of Nonlinear Parameters." *SIAM Journal on Applied Mathematics* 11 (2): 431–41.

Nash, John C., and Arkajyoti Bhattacharjee. 2024. "A Comparison of R Tools for Nonlinear Least Squares Modeling." *The R Journal* 15: 198–215. https://doi.org/10.32614/RJ-2023-091.