Redefining the deviance objective for generalised linear models

Abstract of the Norwich discussion

[Institute and Faculty of Actuaries, 6 June 2011]

Contact

Tony Lovick, F.I.A., E-mail: tony.lovick@towerswatson.com

This abstract relates to the following paper: Lovick, A.C. and Lee, P.K.W. Redefining the deviance objective for generalised linear models. *British Actuarial Journal*, doi: 10.1017/S1357321712000190

The Chair (**Mr D. A. Storman, F.I.A.**)**:** I should like to extend a warm welcome to everyone: thank you for coming along.

The Chair then introduced Mr Lovick in a manner similar to his introduction at the London meeting held earlier on 28 March 2011.

Mr A. C. Lovick, F.I.A. (*introducing the paper*)**:** I am going to cover the motivation for the method, the concepts around what the case deleted deviance is, and some examples. I am then going to try and apply that to the noise reduction method. We have then prepared a number of case studies using real general insurance examples. At the end, I will touch on other examples of how to use the method.

Mr Lovick presented the key findings of the paper in a similar manner to his and his co-author's introductory remarks at the earlier meeting in London.

Dr G. Janacek (*visitor, University of East Anglia, opening the discussion*)**:** I would argue that statistics has revolutionised modelling. However, there is a problem. Traditional statistics is designed experiments. All of statistical inference is based on designed experiments. If you do not collect your data in the right way, then the statistics are likely to be meaningless.

This is a problem which raises all sorts of difficulties. I was brought up in time series, and every time series analyst knows that statistical predictions are wildly over-confident – without exception.

If data is collected in the odd way presented to the authors and they have no real statistical underpinning, what should be done? This is where the paper is very interesting. It seems there are two ways of attacking this very difficult problem.

First, there is the generalised linear mixed model approach. This is where we add random effects to the model predictors; essentially, the model becomes a conditional model. This idea is well threshed out in statistical circles. It is computationally expensive. I am not entirely convinced that this is the way to go, but this would be one possibility, completely different to that set out in the paper.

538

The second is the approach the authors have taken, which is the leave one out, case delete, cross-validation approach.

In the 1930s, people started suggesting a "hold out sample". You fit your data on a training set and then you try it on your hold out sample. This will give you an idea of how good your models are, and you are in some way escaping the statistical straitjacket. This was used intermittently until the rise of the computer and people started talking about cross-validation or K-fold cross-validation. The idea is very simple. You take your data set, you take K observations out of it, fit your model and then you take another different K observations, and fit your model. You do this until you have exhausted all the data. You can use the discrepancy between your K hold out observations and the model as a measure of goodness, or deviation, in some way.

The cheapest, and probably the most popular, version is leave one out cross-validation, which brings us back to the authors' approach. Here we fit all the data but one point, we then compare our model with the leave out point and we repeat this over all the possible leave out points. It was popular in statistics because if you do classical regression, it is easy. The algebra allows you simplification. You do not have to do the computation over and over again, n times. You can do it all in algebra and it works out beautifully. You will have come across the statistics which have been derived. These map very simply to the current case, but need quite a lot of algebra. You get most of it in the paper but the authors leave out the details.

I have some questions for the authors. They have something they call a 'pattern', which is the difference between case deleted deviances. They have something called the 'noise', which is the difference between the deviances and the pattern. They put these together in a linear way to come up with this value. If you say pattern plus alpha noise, the question is: what is alpha? The authors say it is -5.

I concede that you can do this by tuning it; but I would be interested to know how sensitive the problem is to changes in this parameter. In my world of time series analysis, and probably in the machine learning world, there is general agreement that leave one out is bad in the sense that it has maximum variance and really you ought to leave out more than one. People tend to have religious wars about whether it should be 10 or 50. The argument is that as you increase K, the number left out, the variance in your parameter estimates drop but the bias increases. So there is a bias/variance trade-off. I would be interested to hear the authors' view. I realise that the algebra gets much harder but I wonder whether it is possible to move in the direction of using more than one.

Also, I do not understand, if you are adding terms to your model, is there a difference between stepping up and stepping down? I do not think there would be but I do not have a feel for this – if you do leave one out whether you should build your model up or reduce your model. An alternative to this would be to use a bootstrap. I wondered if the authors had thought about using a bootstrap and had rejected the idea or had seen some difficulties that were not immediately apparent.

Mr Lovick *(responding)*: In reply to the first question about why -5 and how sensitive is it, we have not had an awful lot of time to experiment with different values so more work is suggested on that one. We also have not done the theoretical analysis that it should be 5 in the first place. We hope to do this to underpin the experimental evidence with something theoretical as well.

The second question was why not leave out more than one? The reason we decided on the hold out of one, originally was simply the idea that there would be a unique way you could do that.

We aimed to do something that was reproducible each time. So if you hit the 'fit' button and release a set of results, you do not want to get a different answer the next time you do it. So, choosing 10 points to hold out randomly giving different answers (probably similar to your bootstrap method), which requires you to perform enough simulations to make sure that you have an accurate answer each time. Also, systematically going through the data, taking the first 10 points, leaving them out, then the next 10 and the next 10, often using data sorted in some way, normally by time, age or something else, could introduce more bias into the data. So you could not do it by a methodical way of going through the data – it would have to be random in some sense otherwise you would be liable to get the wrong answers.

Stepping up and stepping down is similar to leaving out more than one data point at once. The approximation we give in the paper, case deleted estimate, depends on the variance covariance matrix of the model fit on the whole data. I am a proponent of starting with a full model and then working backwards, because that gives you your full variance covariance matrix to consider at the start.

Also, if you leave out one data point then the variance covariance matrix is valid for calculating how that estimate will change. So that will give you a good approximation, about 99.9% accurate, which we saw earlier.

If you start leaving out more than one data point, then the matrix itself changes. You have got something like a correlation between data points to be considered as well. You might have quadratic effects in there as well. In a sense, the approximation we are using is the first term of a linear expansion. If you leave out more than one data point, you need to start including the quadratic and cubic terms as well.

Mr C. G. Bolton, F.I.A. *(closing the discussion)*: It was good that we had some real examples. As a profession, the results of our work often has an impact on business, shareholder capital, insurance premiums, jobs, and so on. Our professional duty is to ensure that our work is done well, and often we are addressing more than an interesting problem.

I am cautioning us to ensure when we make a conclusion from our results that we make the right conclusion. There is a lot at stake.

Statistics too has developed over the last period by being connected with the challenges of the day and how we try to solve them. Dr. Janacek made the point about how it has transformed modelling.

It was interesting to hear from both Dr. Janacek and Mr Lovick that the people who appear to have least faith in their statistical models are the statisticians themselves. They are compelled to advise caution in their use. Again, it comes back to our professional point. Just because the model, the spreadsheet, appears to say something, it does not necessarily mean it is so. We should add our own wisdom and check before we do anything.

It is a challenge for us to make sure that we are true to ourselves when we do that. It is very easy to post-justify – you get something out of the model and you imagine a world where that is indicative and is an underlying effect. To challenge ourselves, we must ask whether we would have thought that before the model suggested it. Indeed, in some cases, we might remember ourselves arguing the exact opposite when the data was pointing in a different way.

540

It is incumbent upon us to look at results with a critical eye and make sure that we have not inadvertently, and without consciously knowing it, changed our view. That particularly is necessary because adding an additional point or an additional feature to our models always seems to get a payback. Mr Lovick drew our attention to the fact that almost every new rating factor that you put in explains some of the residuals and therefore could potentially be true. Again, it is incumbent upon us to make sure when we are acting in our professional capacity to look at it again with a critical eye.

Dr Janacek pointed out that just because it makes for a good paper does not necessarily mean it is ready for the big world. Clearly, there is an ongoing debate of how we can refine aspects of the paper before us to be more consistent with academic tradition. I also notice that, however much Mr Lovick denies it, because he is using some case-deleted trick, he does have a secret hold out sample. It is always worth pressing him on that and making sure that he comes up with all the goods.

I am grateful to the general insurance part of the Actuarial Profession for their statistical modelling. They have really led the way in modelling over the last few years. It has been incredibly tough in the motor market. You have had to work your models really hard. In life assurance, we have started the first part of that journey with things like postcode and some of the marital status data we have, and using some very light generalised linear modelling. This has proved enormously commercially effective in applying it to businesses like annuities.

For Aviva, at least, it has enabled them to grow the business on the back of the techniques that GI have developed in motor and also household. It does not need me to tell you that analysing postcode with all the 1.6 to 1.7 million full postcodes, is not as easy as treating all the postcodes differently. You have to attach all the factors to that to build up your secondary statistics to use it. We applied it to our business to great effect.

The Chair: It is nice to finish on a positive note. If anyone else in the audience would like to raise anything, this is your last opportunity.

In that case, I should like to give my personal thanks, and I am sure you will join me, to Mr Lovick, Dr. Janacek and Mr Bolton for their excellent contributions.