
On the Accuracy of GLMs

Summary Document

March 2009

by:
Dr. Paul Beinat



For a very long time the hands-on practice of statistics was based on parametric approaches using model-based methods and books of tables. These methods became well known and widely practiced with considerable amounts of intellectual capital and experience being developed around, and on top of, each of the methods. Perhaps the most widely known and used is the linear model. In the international P&C (Property and Casualty) industry the most powerful parametric model in use for rate setting today is the Generalized Linear Model (GLM). The GLM originally appeared in the early 1970s and predates the introduction of the personal computer; it was initially a pencil and paper method. Since the 1970s the GLM has gradually gained wide acceptance and is today seen as the benchmark for P&C classification ratemaking.

Over the past 30 years a number of methods that are computationally intensive have been discovered that would not have been possible without the advances in computers and computer science¹. Many of these methods apply brute empiricism in order to get a result and in turn do not produce their own unique set of validation statistics. There is a tendency of any model to overfit on the data on which it is trained. In order to overcome this overfitting the computer scientists opted for validation. Generally validation is performed by splitting the data into 2 groups: one for training the model and the other to check or validate the result. For computer scientists validation is the true test of the effectiveness of the model. The question that is natural for any scientist to ask given this background is – how to validate GLM produced model results?

For this exercise 2 relatively small personal lines portfolios were selected, one Auto collision and the other Homeowners, portfolios of this size being the most commonly occurring. The data for both portfolios was split into training and validation samples. GLMs were fitted to the training data for claim frequency; claim frequency generally provides a much better fit than claim size GLM models. Both models were run through a GLM optimization process that tests every possible combination of supplied variables. The winning model selected in both cases was the model with the best statistical inference, i.e. the model with the best model statistics indicating fit. The 2 models were then applied to their respective sets of validation data and the results compared. Note: In order to be able to apply a universal measure of deviance 1 minus the deviance of the model divided by the deviance of the null model was used ($I = 1 - \frac{d_{model}}{d_{null}}$). This measure shows the relative improvement of any model over the null model applied to the same data and, therefore, enables comparisons of deviance between the training and validation models. The results:

¹ "In a world in which the price of calculation continues to decrease rapidly, but the price of theorem proving continues to hold steady or increase, elementary economics indicates that we ought to spend a larger and larger fraction of our time on calculation", Professor John W. Tukey. http://en.wikipedia.org/wiki/John_Tukey

- Auto Collision Portfolio

3 years data; 70% training, 30% validation, randomly split.

	Training	Validation
Deviance	6356.6198	5007.7991
Pearson Stat	5845503.9905	8822293.6783
AIC	0.2081	0.2869
BIC	-758078.9118	-363860.2653
Efron Pseudo-R ²	0.0145	0.0043
McFadden index	0.0350	0.0163

Table 1

The training deviance improvement, $I = 7.5\%$.

In deviance terms the model does not do much better than just using the average.

The validation deviance improvement, $I = 3.2\%$.

The results from the deviance improvement measure are consistent with the Efron measure.

Effectively less than half of the signal is present in validation with the conclusion to be drawn from this being – the GLM model based on the training data overfits even when it shows excellent model statistics (no validation). Note: this is consistent with the Bayesian view of models fitted using maximum likelihood, they overfit.

- Homeowners Portfolio

Modeling on one year, training one the next year (real life scenario).

	Training	Validation
Deviance	2286.8363	4971.1931
Pearson Stat	21.5525	15.6539
AIC	0.0003	0.0004
BIC	-321306723.6643	-324074651.0975
Efron Pseudo-R ²	0.4449	-0.6298
McFadden index	0.27	-0.2382

Table 2

On the validation data the Efron R² is very negative indicating that the mean would do much better than the GLM model. The deviance improvement measure exhibits the same kind of deterioration, model – $I = 49\%$ and validation – $I = -40\%$. From a model perspective it is not good, but the reverse in the signal is driven by one variable which was disclosed by the validation process. Re-running the model with a random 70/30% data-split delivers a far more satisfactory result. The Efron R² is 0.7623 for training and 0.5667 for validation. This is a good model, but with an obvious element of overfitting.

- Why not use Gini Coefficient as a measure of fit?

The Gini Coefficient is a measure of lift and is not a measure of fit. For example the first version of the Homeowners' model did not fit well (Efron R² of 0.4449 for training and -0.6298 for validation); however despite this very poor fit the Gini Coefficients were 0.2930 and 0.23453

respectively. The fit was poor, but the Gini Coefficient was good. The Gini Coefficient simply shows that the lift was comparable between the training and validation. It is not a measure of fit.

- Simple Statistical Methods

General Iterative Algorithms² are a generalization of the older minimum bias models; they represent a more flexible approach to ratemaking than GLMs. Three parameters control how each GIA fits the training data – these are referred to as P, Q and K. Certain values of the variables correspond to a log link and Poisson error structure, or a log link and a gamma error structure. For instance when all three variables are equal to 1 this corresponds to the usual Poisson GLM, while values of 1, 0 and 1 respectively correspond to the gamma GLM. However, many other combinations of these three variables are possible and do not correspond to any commonly used GLM models.

The GIAs come with their own set of statistics that are used to rank the results from the multitude runs required to exhaust all of the combinations of P, Q and K values. The GIA measures are general and can be simply applied to the GLM results thereby enabling elementary comparison of GLM versus GIA results. The general GIA measurements are:

- Weighted Absolute Bias (WAB)
- Weighted Absolute Percentage Bias (WAPB)
- Weighted Pearson Chi-squared (WChi)
- Combined absolute bias and Pearson chi-squared statistic (WABWChi)

Auto Collision Portfolio. Running the Auto Collision portfolio referenced previously using the GIA on identical training and validation samples: If we determine the best model based on the validation WAB then the Poisson GLM is 2nd best result. If we rate performance on WAPB then the GLM is 428th. Based on WChi the GLM is 394th. While on WABWChi it ranks 212th.

Homeowners Portfolio. Running the Homeowners portfolio in a similar manner and comparing the best GLM result to the GIA result: Using WAB the Poisson GLM is 22nd, using WAPB then the GLM is 20th. On WChi the GLM is 7th. While on WABWChi the GLM result is ranked is 8th. What was a very good GLM result performs quite poorly by comparison to the GIA-result.

In addition to the experiments listed above similar work has been performed on some very large portfolios that have had the GLM applied by: a top 5 personal lines insurer, by different groups of consulting actuaries using the most popular commercially available GLM software, and the similar signals were exposed in each instance. This type of experiment has been conducted so many times over the past 2 years that it is obvious that the consistent results achieved are an attribute of the GLM method not the data or of the analyst fitting the method.

² Fu L and Wu P General Iteration Algorithm for Classification Ratemaking [Journal] // Variance. - [s.l.] : CAS. - 02 : Vol. 01.

The conclusions that can be drawn the last 2 years experience and experiments:

- Model statistics can be misleading. Good sets of statistics produced by a GLM on training data do not mean that the GLM results will validate well.
- GLM assumptions regarding error structure and link function are not optimal. The GIAs with their iteration over a large number of options clearly show this.
- Some of the issues with the GLM fit relate to the use of maximum likelihood to produce a solution. The log likelihood surface is very shallow with the maximum on this flat surface dependent on idiosyncrasies in the data. As much as there are methods and heuristic knowledge that attempt to address this issue the problem remains and can be seen in the deviance fluctuations when running stepwise or optimized GLMs. Maximum likelihood is not immune to squared error style problems.
- GLM Models overfit the data and the extent of this can be seen by use of validation data. Bayesian mathematicians believe that they have a proof that the maximum likelihood method generally overfit the data.

This paper is a summary of a more detailed scientific paper on this topic. This paper runs to approximately 40 pages.

A complete paper copy of this paper can be obtained by contacting:

Sales Department

803-726-7214

salesinfo@eeanalytics.com