GLMs – the Good, the Bad, and the Ugly

Midwest Actuarial Forum

23 March 2009

Christopher Cooksey, FCAS, MAAA

EagleEye Analytics

Agenda

1. A Brief History of GLMs
2. The Good – what GLMs do well
3. The Bad – what GLMs don't do well
4. The Ugly – what GLMs can't do
5. Solutions

Section 1

GLM History

A Brief History of GLMs

- Formulated by Nelder and Wedderburn in 1972.
- First edition of McCullagh/Nelder book on GLMs in 1983.
- One of the first examples of use in insurance was "Statistical Motor Rating: making effective use of your data" by Brockman and Wright in 1992.
- "Practitioner's Guide to Generalized Linear Models" written in 2007.

Section 2

The Good – what GLMs do well

The Good – what GLMs do well

- There is an established and understood literature.
- There is increasing DOI acceptance.
- There are readily available software solutions.
- GLMs extrapolate over predictor levels with little or no data.
- GLMs provide easily calculated relativities to use as a rate classification system.
- GLMs clearly find significant signal in insurance data.

The Good – what GLMs do well

- GLMs are parametric and come with all the advantages of parametric approaches.
  - By assuming you know the form of the "noise" you can do statistical inference to evaluate predictors.
  - You can also provide confidence intervals to communicate the inherent uncertainty in the output.
  - Parametric approaches are very accurate when the assumptions hold.

Section 3

The Bad – what GLMs don't do well

The Bad – what GLMs don't do well

- The assumptions underlying GLMs may not hold.
- Investigating this issue takes time, as do corrections to the basic assumptions (if necessary).
- Issues include…
  - Appropriateness of the link function
  - Appropriateness of the error function
  - Predictiveness of the model

The Bad – what GLMs don't do well

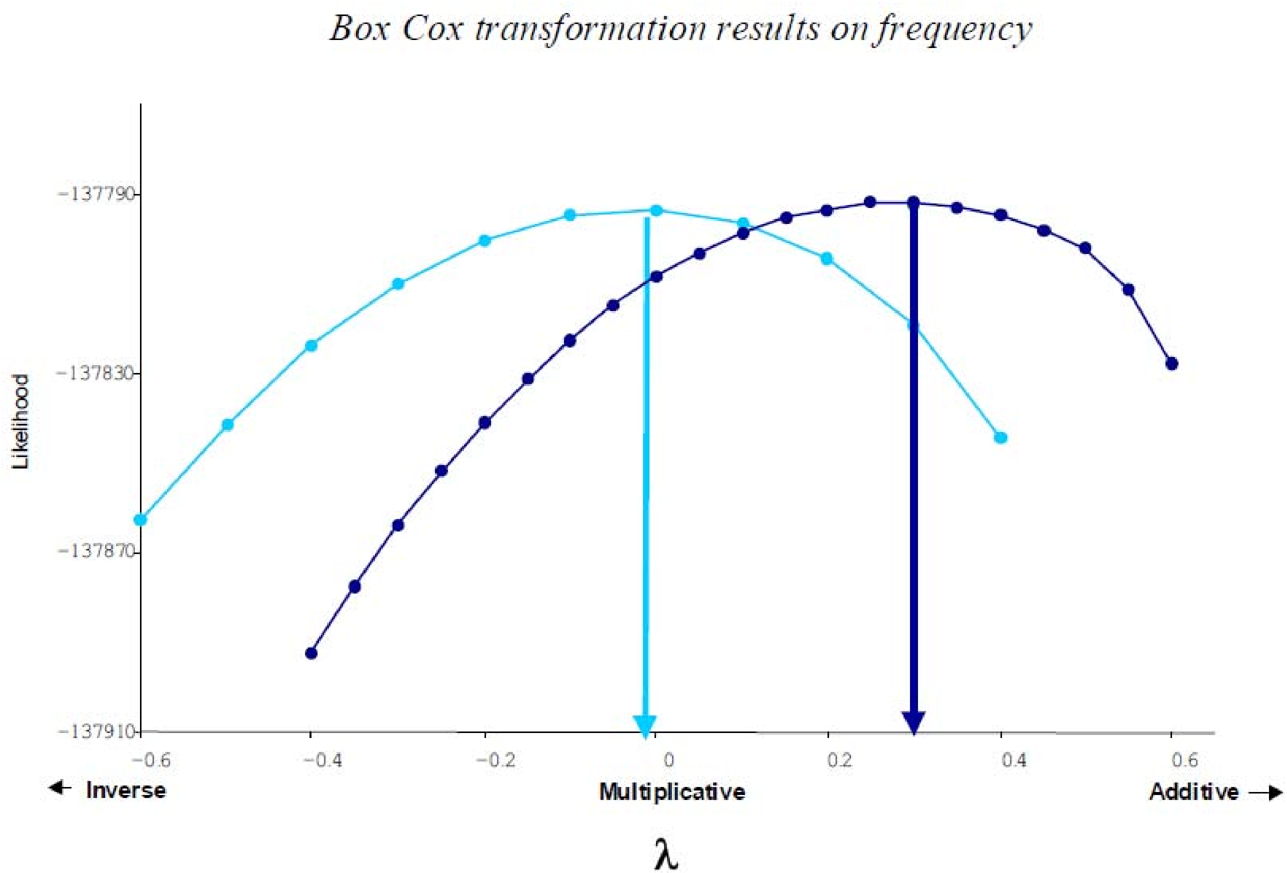
One assumption is that the log link works well for insurance data.

- This can be tested with a Box Cox Transformation (an example of this can be found in the "Practitioner's Guide").
- Use the following link function.

$$g(x) = (x^{\lambda} - 1)/\lambda \qquad \text{when } \lambda \neq 0$$

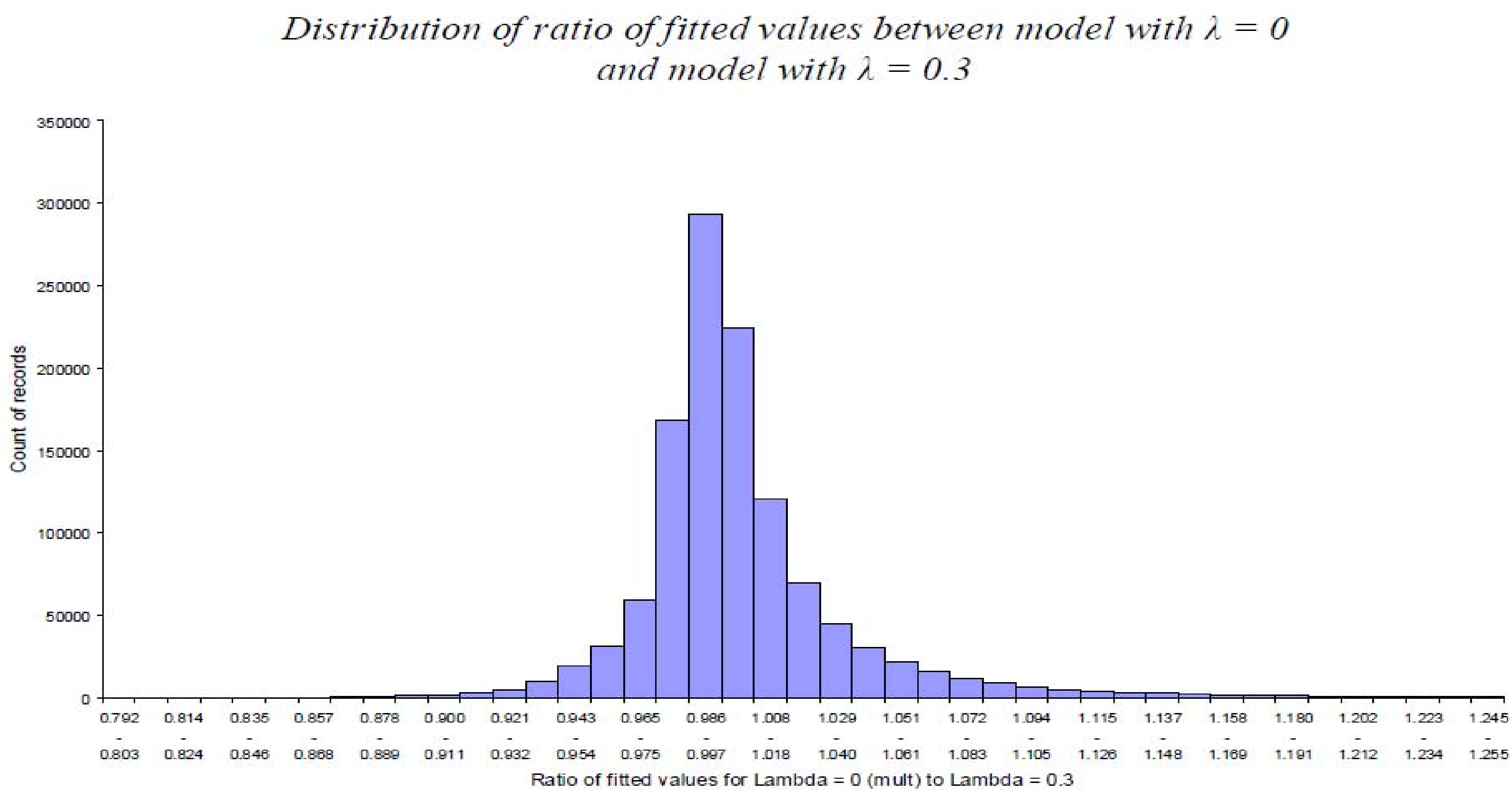
$$g(x) = \ln(x) \qquad \text{when } \lambda = 0$$

GLMs – the Good, the Bad, and the Ugly



Taken from "A Practitioner's Guide to Generalized Linear Models", Third Edition, page 59.

GLMs – the Good, the Bad, and the Ugly



Taken from "A Practitioner's Guide to Generalized Linear Models", Third Edition, page 60.

The Bad – what GLMs don't do well

One assumption is that the log link works well for insurance data.

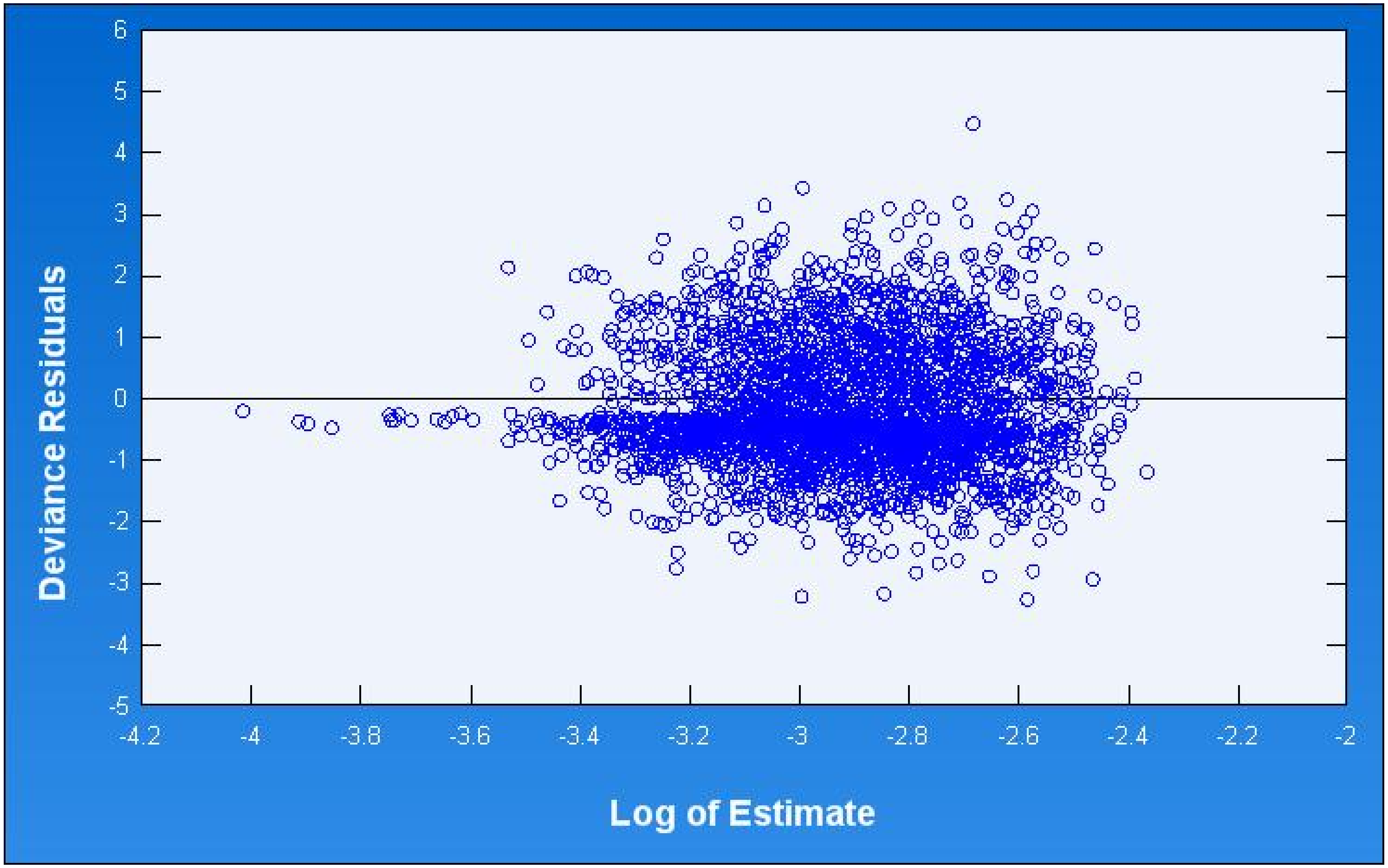
- There is no fundamental reason that real insurance data must be multiplicative in nature.
- Usually, a multiplicative model is more appropriate than other model forms.
- Consequently, multiplicative models are used. This is usually counted as a minor distortion.

The Bad – what GLMs don't do well

A second assumption is that the typical error functions (Poisson and gamma) work well for insurance data.

- This can be tested by looking at the residuals.
- Many things can be done to correct for patterns in residuals, but rarely, if ever, do you have perfectly homogeneous residuals.
- Sometimes you can correct for known distortions (zero-inflated Poisson, for example).
- These issues are usually counted as minor distortions.

GLMs – the Good, the Bad, and the Ugly



The Bad – what GLMs don't do well

The predictiveness of the model is an additional assumption that sometimes isn't considered.

- Most of our time is spent on significance testing predictors against the training data.
- This can lead to the erroneous assumption that if a model tests well on training data that is will also validate well on hold-out data.
- Significance testing alone tends to overfit models.
- Not a fundamental issue with GLMs, but a common problem instead.

The Bad – what GLMs don't do well

The final category of issues with GLMs revolves around the time and effort involved in doing them well.

- While all advanced modeling techniques require knowledgeable practitioners, GLMs involve an extensive modeling process.
- Modelers fluent with the details are required to "shepherd" the modeling process itself.
- Also, the relatively numerous and statistically strong assumptions of GLMs require evaluation throughout.

Section 4

The Ugly – what GLMs can't do

The Ugly – what GLMs can't do

GLM model risk can be mitigated but not removed.

- There is no theoretical reason that any given error function should fit insurance data precisely.
- There is no theoretical reason that signal in insurance data must be related to multiplicative predictors.
- There is always some risk that the imperfections of the model assumptions will substantively impact results.

The Ugly – what GLMs can't do

GLMs simply do not provide a system for finding all of the relevant interactions.

- It is not practically possible to test through trial and error all possible combinations of three-way interactions, let alone interactions involving four, five or more predictors.
- The key question is: Do three, four, five-way interactions exist in some way that meaningfully impacts results?
  - Our research says they do. An example will follow.

The Ugly – what GLMs can't do

Another problem with interactions is that GLMs are not formulated to find local interactions.

- GLMs use global interactions – the interaction between all levels of two predictors.
- Once this interaction is included, it is possible to note relevant portions and to smooth over irrelevant portions, thus creating local interactions between only certain levels of each predictor.
- This process is only practical for simple interactions.

Section 5

Solutions

Solutions

Keeping in mind a realistic view of GLMs, there are at least three possible responses.

1. Continue to rely solely on GLMs
2. Abandon GLMs for some other alternative
3. Find some supplement to cover for GLMs' weaknesses

Solutions

If you stick with GLMs, remember the difficulties…

1. GLMs are parametric. Model assumptions impact the results.
   - Make sure you test the assumptions and consider alternatives to the typical Poisson/frequency and gamma/severity combinations.
2. GLMs provide no good way to explore the universe of possible interactions.
   - Make sure you set aside time to find these. Use intuition and scan your competitors for options. Also look for where your model is out of balance – where observed losses are not close to predicted losses for significant segments of the book of business.

Solutions

If you abandon GLMs, what else is there?

- Data mining techniques
- Minimum bias
- General Iteration Algorithms (Fu, Wu, 2007)
- Something else???

Solutions

A third approach is to find a supplement to GLMs. Again, consider the difficulties…

1. GLMs are parametric. Model assumptions impact the results.
2. GLMs have no good way to explore the universe of possible interactions.

All you need to find is a nonparametric, nonlinear approach which quickly finds relevant local interactions.

Solutions

What possible candidates exist for accomplishing this? There are many nonparametric approaches and other tools to be found in the fields of data mining and machine learning…

- Neural networks
- MARS
- Decision trees
- CART, CHAID
- Random forests
- Polynomial networks
- Principle components
- Kernels
- Bagging
- Boosting
- Bootstrapping & resampling
- Activity mining

Solutions

Some issues in developing a solution include…

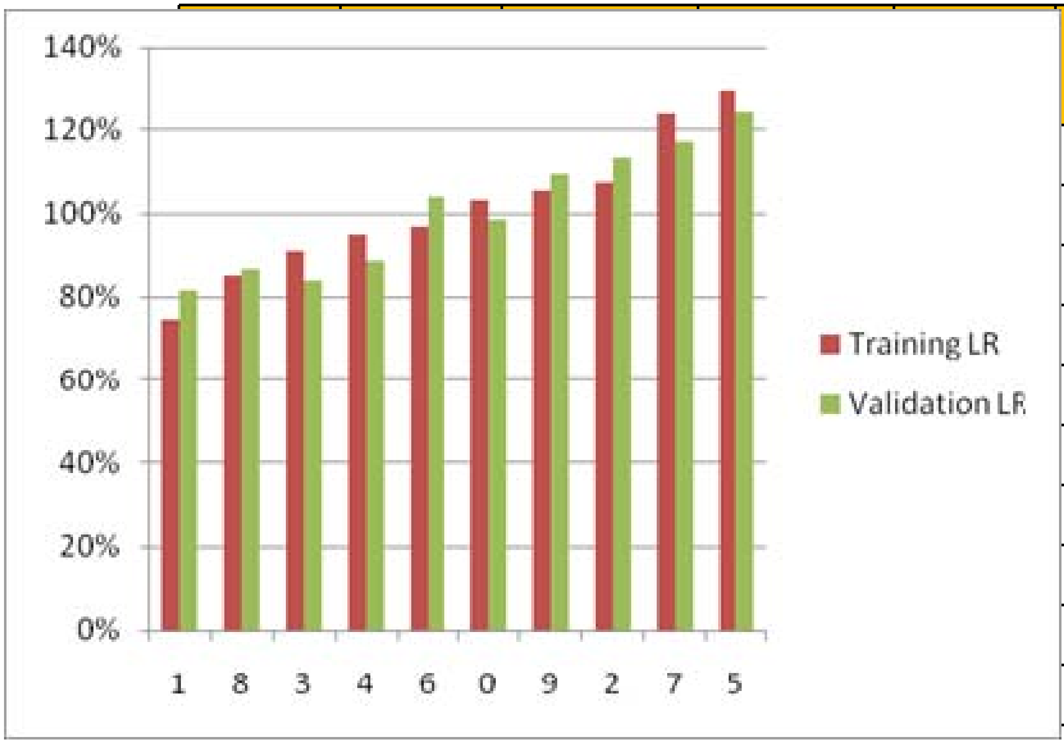
- Getting the technical expertise in nonparametric solutions.
- One-size-fits-all data mining methods have shown moderate performance on insurance-specific data.
- Better results are found by ensembling multiple methods.
- Nonparametric methods tend to be greedy – significant risk of overfitting.

Example

- Medium size auto portfolio
- Pure premiums set by consulting actuaries using GLM tools on 2 years of data
- Used machine learning inductive techniques on the first year of data (training dataset) to search the residuals of the GLMs
- The second year of data was used as validation
- Used policy attributes only

[Example pulled from "Machine Learning vs. Traditional Methods – Summary Document" by Dr. Paul Beinat. Paper can be found at the EagleEye Analytics website – Resources, Papers.]

Example



Training LR	Validation LR
75%	82%
85%	87%
91%	84%
95%	88%
97%	104%
103%	98%
105%	109%
107%	113%
124%	117%
129%	124%

Example

- Correlation between training and validation loss ratios: 0.93
- Lift is similar between the two datasets:
  - There is a similar difference between maximum and minimum
  - There is a similar exposure-weighted standard deviation of the loss ratios: Training is 16.4%; Validation is 14.5%
- Significant improvement in fit on the validation data when

	Deviance	Squared error	Chi squared error
GLM Premiums	34.15388	1463.838	5.47708
Estimated Premiums	15.02064	243.8955	0.946702

Section 6

Questions?

Contact Info

Christopher Cooksey, FCAS, MAAA

EagleEye Analytics

ccooksey@eeanalytics.com

www.eeanalytics.com