autoPricing: An R package for automated GLM based actuarial pricing

Chibisi Chima-Okereke

cchima-okereke@mango-solutions.com

Introduction

Automation of actuarial pricing

- Outlining the current barriers against automating actuarial pricing
- Proposed scheme
- Rebalancing rating factors
- Selection criteria
- Defining consistency
- Forward/backward stepwise process
- autoPricing package
- Conclusion

Introduction

Preliminaries

MANGOSOLUTIONS
data analysis that delivers

Focus & Assumptions

- This presentation is not about Motor pricing!
- The presentation focuses on the difference between this and standard actuarial approaches rather than the features of the package
- Focusing on GLM based pricing
- Only models with categorical explanatory variables will be considered
- Rating Factor == Explanatory Variable
- Null Model ==> Model with Year category

Introduction

Preliminaries

MANGOSOLUTIONS
data analysis that delivers

Why should we automate GLM pricing?

- The current process is very labour intensive
- Labour == Expense
- Current process is manual
- Manual == Human Error Potential ==> Large amounts of checking
- Manual == Difficult to integrate with other modern systems
- Because actuarial analysts have better things to do with their time – well don't you?

Outlining the problem

Automating actuarial pricing

MANGOSOLUTIONS
data analysis that delivers

Typical workflow in GLM pricing exercise

- Data Preparation – policy and claims table as text files for input into pricing software
- Data Importer – internal pricing software data preparation
- Optional data analyser
- Manual stepwise analysis for risk and severity model
- Model combiner for risk and severity model
- Model output

Outlining the problem

Automating actuarial pricing

MANGOSOLUTIONS
data analysis that delivers

Typical manual stepwise analysis after ordering variables:

- Select explanatory variable or rating factor
- Fit model
- If significant fit year interaction else adjust variable and try again
- If not significant rebalance or reject variable
- If consistent accept variable
- If not consistent reject variable
- Next variable

Outlining the problem

Automating actuarial pricing

MANGOSOLUTIONS
data analysis that delivers

Challenges and issues with automating insurance pricing of personal lines

- Current stepwise and combinatorial selection methods have not incorporated actuarial alterations to the model between selection steps
  - Definition of consistency is currently qualitative and dependent on individual observer to some extent
  - Categorical variables are altered and aggregated subject to significance of factor levels and the logic of their aggregation
- Current actuarial pricing software has not incorporated the above methods

Proposed Scheme

Automation of actuarial pricing

MANGOSOLUTIONS
data analysis that delivers

The **autoPricing** package in R using the following scheme to resolve these problems

- Use the current framework for stepwise model selection
  - Forward selection
  - Backward elimination
- Rebalance the variable using a combination of pairwise Tukey test and logical mapping table supplied by the user
- Quantify the consistency of the rating factor so that we select on it using a threshold
- Decision to use information criterion since the concept is simple, straightforward and more flexible than significance.

Rebalancing variables

If your variable is not significant

MANGOSOLUTIONS
data analysis that delivers

Rating factor mapping table example of age

Initial Factor Levels	Aggregatable Levels
≤ 18	Age_1
19 – 25	Age_1
26 – 35	Age_2
36 - 55	Age_2
56 – 65	Age_2
> 65	Age_3

Pairwise Tukey Test

Automation of actuarial pricing

MANGO SOLUTIONS
data analysis that delivers

The pairwise Tukey HSD (Honestly Significant Difference) Test for coefficient β_i

$$H_0: \beta_{i_A} = \beta_{i_B}; \qquad H_A: \beta_{i_A} \neq \beta_{i_B}$$

$$Reject\ H_0\ if\ \left|\beta_{i_A} - \beta_{i_B}\right| > HSD = q_{A,\alpha}\sqrt{\frac{1}{2} MSE\left(\frac{1}{S_a} + \frac{1}{S_{a'}}\right)}$$

Where β_i is the fitted coefficient, HSD is the threshold, S_a is the number of observations in the group a, and $q \sim \mathrm{Student}(\nu)$. Where $\nu = N - A$, N = No. Observations, $A =$ No. Groups, and α is the level of significance.

Selection Methodology

Information Criteria

For a model $\boldsymbol{\Theta_C}$ and the submodel $\boldsymbol{\Theta_S}$, $\boldsymbol{\Theta_C}$ is considered if:

$$IC(\boldsymbol{\Theta_C}) < IC(\boldsymbol{\Theta_S})$$

Where $\boldsymbol{C} \cap \boldsymbol{S} = \boldsymbol{C_R}$ is the rating factor under consideration and IC() is the information criterion operator.

Forward/backward scheme

MANGOSOLUTIONS
data analysis that delivers

Forward addition

- Start with model $\boldsymbol{\Theta_S}$

$$\eta = \alpha_i Year + \zeta$$

- Consider rating factor $\boldsymbol{C_R}$ where $\boldsymbol{C_R} \cup \boldsymbol{S} = \boldsymbol{C}$ adding if $IC(\boldsymbol{\Theta_C}) < IC(\boldsymbol{\Theta_S})$ or else reject and go to next item

Backward elimination

- Start with model $\boldsymbol{\Theta_C}$

$$\eta = \alpha_i Year + \boldsymbol{\beta}[\boldsymbol{RF}] + \zeta$$

Consider rating factor $\boldsymbol{C_R}$ where $\boldsymbol{C_R} \cup \boldsymbol{S} = \boldsymbol{C}$ keeping if $IC(\boldsymbol{\Theta_C}) < IC(\boldsymbol{\Theta_S})$ or else elimination and go to next item

R. R. Hocking, The Analysis and Selection of Variables in Linear Regression, Biometrics, Vol. 32, No. 1. (Mar., 1976), pp. 1-49.

Consistency

What does consistency mean?

That the levels of the rating factor move together across all exposure years

- Divide the rating factors categories into pairs
- For each year, we see if the fitted coefficient moves up, down or no change in the interaction model
- We take the percentage of the direction having the greatest frequency across all the years for each categorical pair
- Do this for all the categorical pairs in the rating factor and take the median

Consistency

Formal definition

Rating factor coefficients $j = 1, 2, 3, \ldots, J$ for J categories in the rating factor (RF), the model is given by:

$$\eta_{ij\boldsymbol{V}} = \log E(Y_{ij\boldsymbol{V}})$$

$$\eta_{ij\boldsymbol{V}} = \mu + \alpha_i Year + \beta_j RF_k + (\alpha\beta)_{ij} Year\colon RF_k + \boldsymbol{\beta}[RF]_{\boldsymbol{V}}^{k} + \zeta$$

$$\eta'_{ij} = \alpha'_i + \beta'_j + (\alpha\beta')_{ij}$$
$$\delta_{ij} = sgn\left(\eta'_{i(j+1)} - \eta'_{ij}\right)$$

Consistency

Formal definition

Matrix δ_{ij} for calculating consistency for Bonus Malus rating factor with 14 levels

Factor Level Pairs

Year	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10-11	11-12	12-13	13-14
2006	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1
2007	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1
2008	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	1
2009	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1
2010	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1
	100%	**100%**	**100%**	**100%**	**100%**	**80%**	**100%**	**100%**	**100%**	**100%**	**100%**	**80%**	**100%**

Percentage Consistency

Consistency

Formal definition

Rating factor coefficients $j = 1, 2, 3, \ldots, J$ for J categories in the rating factor, the model is given by:

$$\forall j \in \{1, \ldots, J-1\} \quad \Gamma_j^l = \frac{1}{I}\sum_{i=1}^{I} \mathbb{I}_{[\delta_{ij}=l]} \quad where \quad l = -1, 0, 1$$

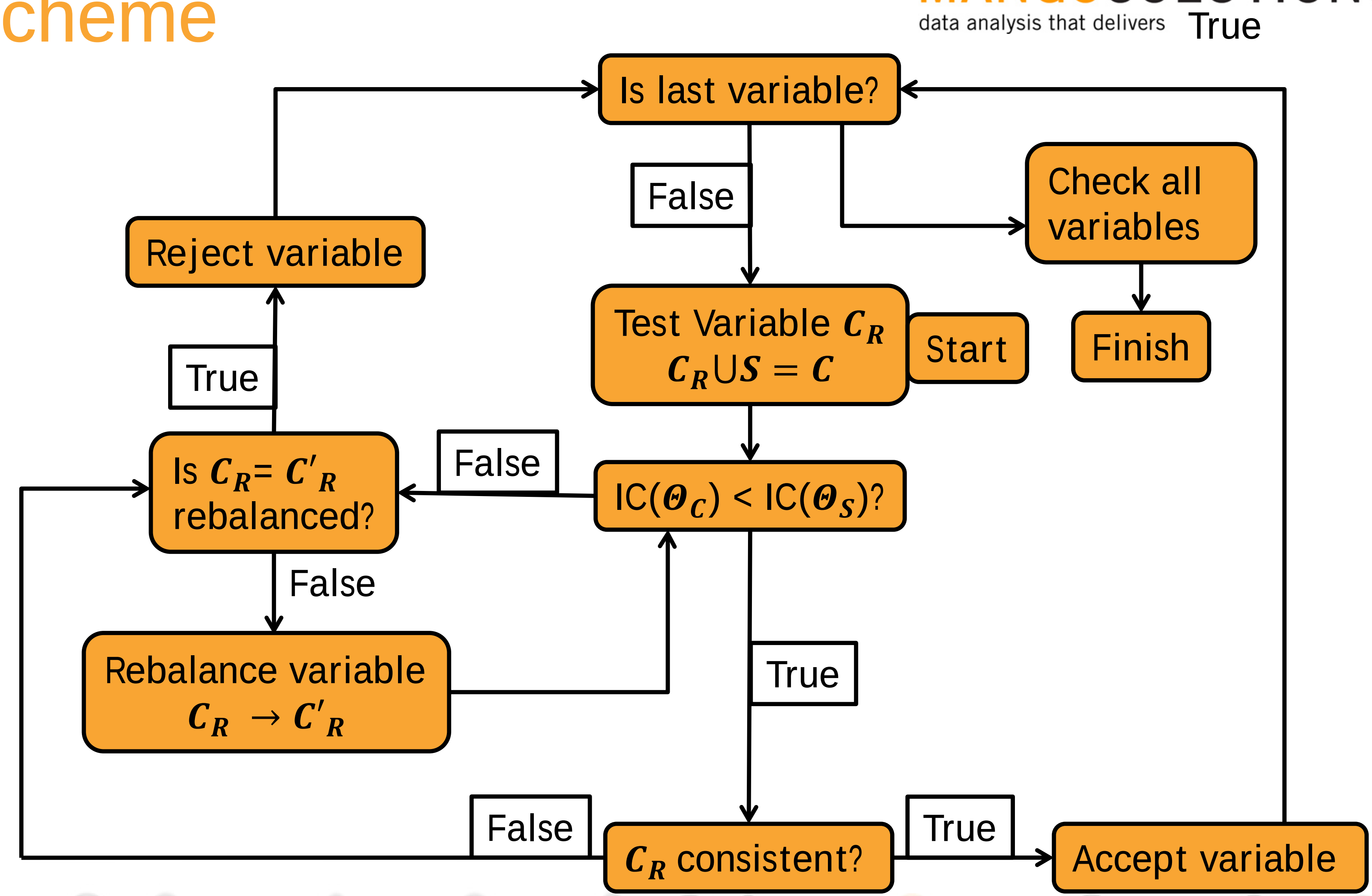
For each year i, the consistency measure over the rating factor

$$\Gamma_j = \max_l \{\Gamma_j^l\} \quad j = 1, 2, 3, \ldots, J-1$$

$median(\Gamma) = \tilde{\Gamma}$ and the consistency threshold is γ

If $\tilde{\Gamma} > \gamma$ the rating factor is consistent

Forward/backward scheme



autoPricing Package

Data shape

Year	Exposure/ Years	Number of Claims	Gross Incurred	Rating Factor
2006	2	1	1144	Group A
2007	3	0	0	Group A
2009	1	0	0	Group B
2010	1	0	0	Group A

Factor | Numeric | Integer | Numeric | Factor | Object Type

Data Source (Simulated): Modern Actuarial Risk Theory Using R: Kaas, Goovaerts, Dhaene, and Denuit.

autoPricing Package

Variable Description

Package is based around a single function **`stepIC()`**

Arguments	Description
`ratingFact`	Character vector denoting the column headers of rating factors in the data table
`countVar`	Character denoting claims count
`sevVar`	Character variable denoting severity variable
`factLevels`	A list of matrices or data frames denoting the mapping of the rating factors
`timeVar`	Variable denoting the year variable
`IC`	Either AIC or BIC
`consistThresh`	Set this to the median of the consistency threshold
`theData`	Dataset that will be used for the analysis
`analysisType`	“frequency” or “severity” flag
`myDistr`	Character string denoting the distribution that will be used
`theLink`	Link function
`exposureName`	Character string denoting the column name of the exposure
`myDocumentTitle`	Character string for the document title
`theAlg`	This is the algorithm to be used either "Forward" or "Backward"

autoPricing Package

Example Code

MANGOSOLUTIONS
data analysis that delivers

Example 1: Executing forward algorithm for Poisson risk model

```
outputModelForwardFreq <- stepIC(ratingFact =
       myRatingFactors, countVar = "NoClaims", sevVar =
       "GrossIncurred", factLevels =
       ratingFactorLevels, timeVar = "Year", IC = "BIC",
       consistThresh = 60, theData = policyTable,
       analysisType = "frequency", myDistr = "poisson",
       theLink = "log", exposureName = "Exposure",
       handicap = 0, myDocumentTitle = "Automated Pricing
       GLM", theAlg = "forward", plotCharts = TRUE)
```

pdf() can be used to write the output graphs to file and sink() can be used to divert the prints to a log file.

autoPricing Package

Summary of GLM

MANGOSOLUTIONS
data analysis that delivers

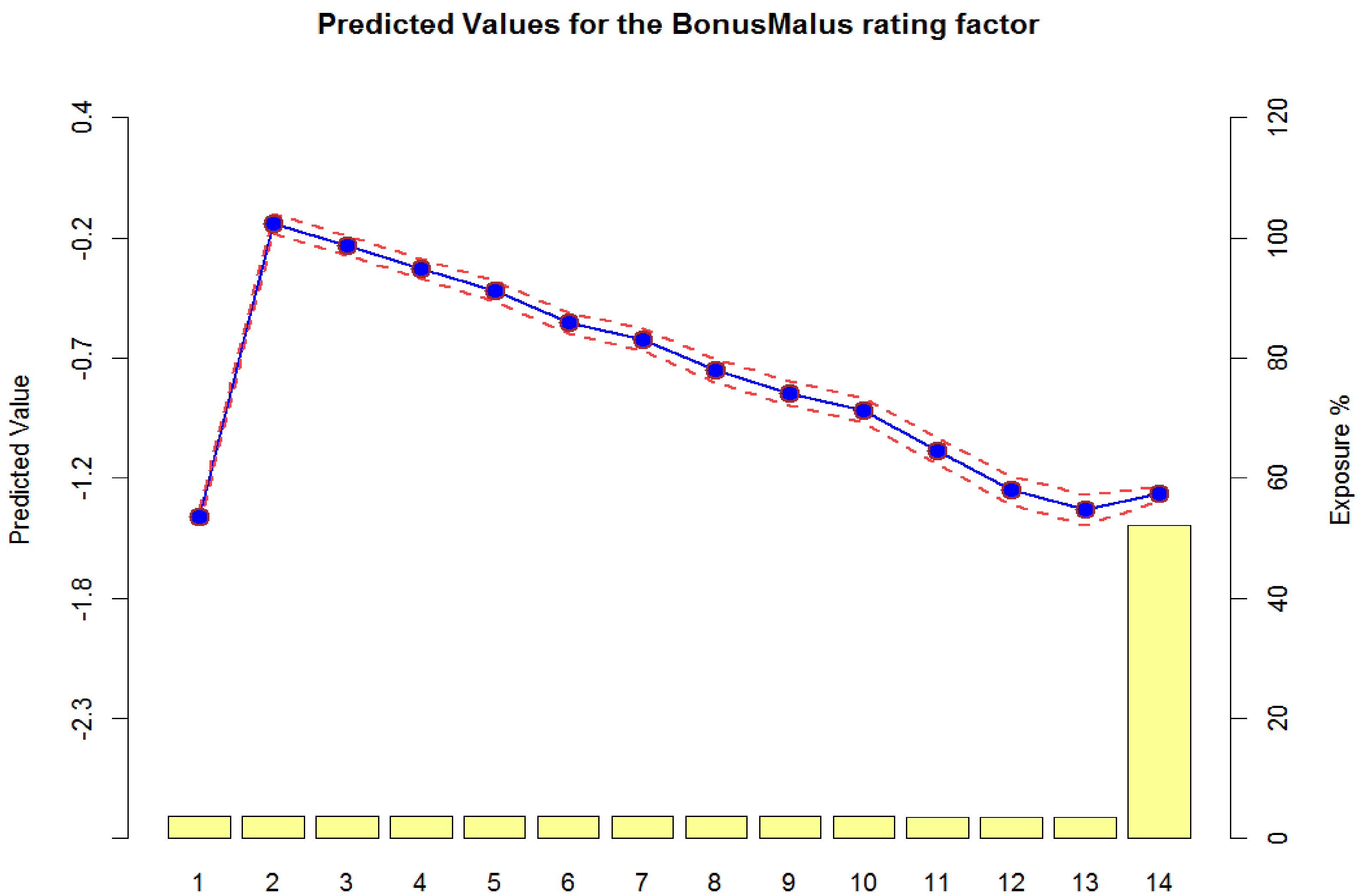
```
> print(summary(outputModelForwardFreq)$coeff, digits = 2)

             Estimate  Std. Error z-value Pr(>|z|)
(Intercept)   -1.3971      0.017  -81.49  0.0e+00
Year2007      -0.0151      0.014   -1.09  2.8e-01
Year2008      -0.0138      0.014   -1.00  3.2e-01
Year2009      -0.0204      0.014   -1.47  1.4e-01
Year2010      -0.0069      0.014   -0.49  6.2e-01
BonusMalus2   -0.0939      0.021   -4.40  1.1e-05
BonusMalus3   -0.1935      0.022   -8.81  1.2e-18
BonusMalus4   -0.2961      0.023  -13.09  3.6e-39
    .             .            .        .        .
    .             .            .        .        .
    .             .            .        .        .
```

autoPricing Package

Predicted Values

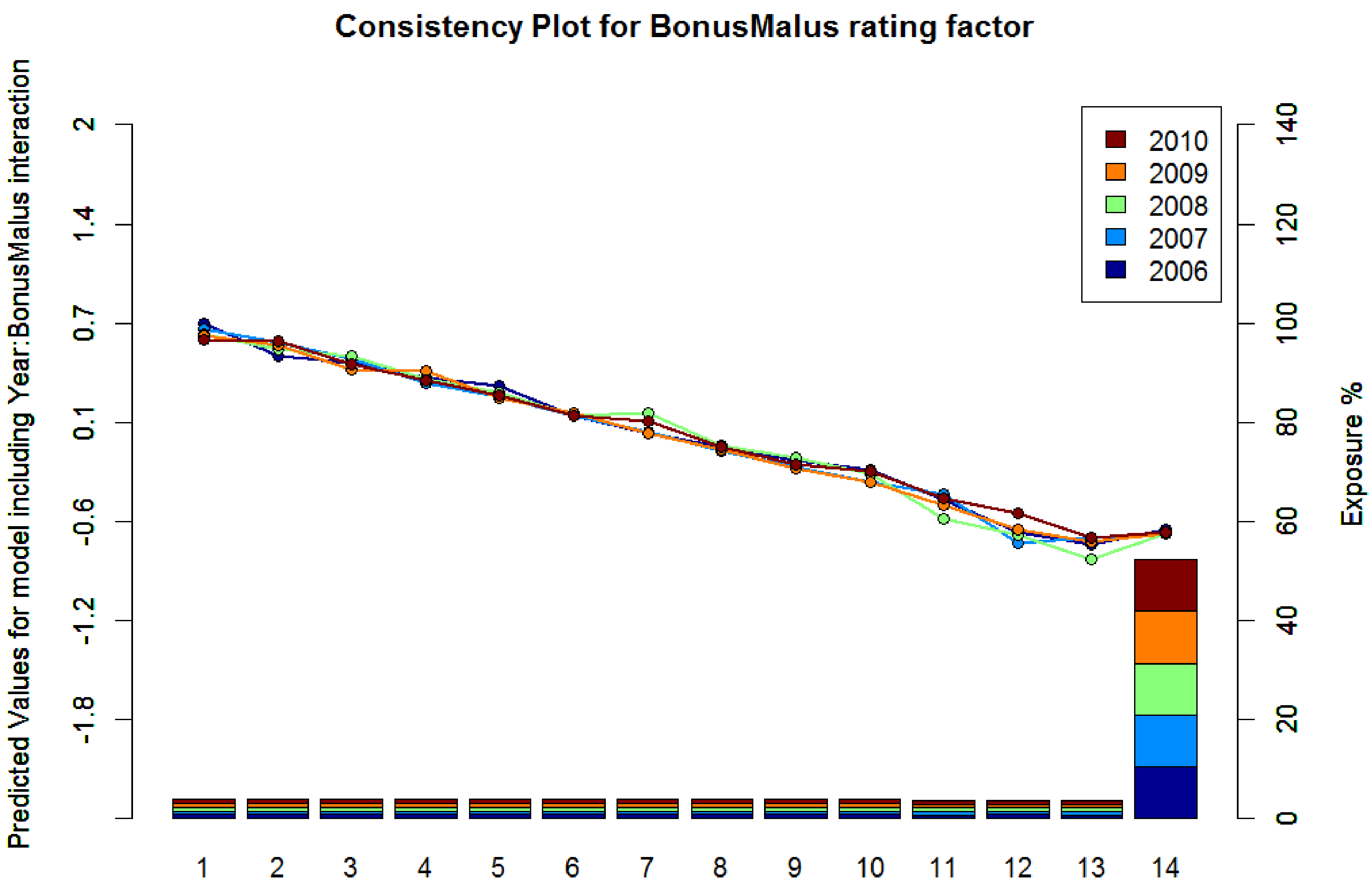
MANGOSOLUTIONS
data analysis that delivers



autoPricing Package

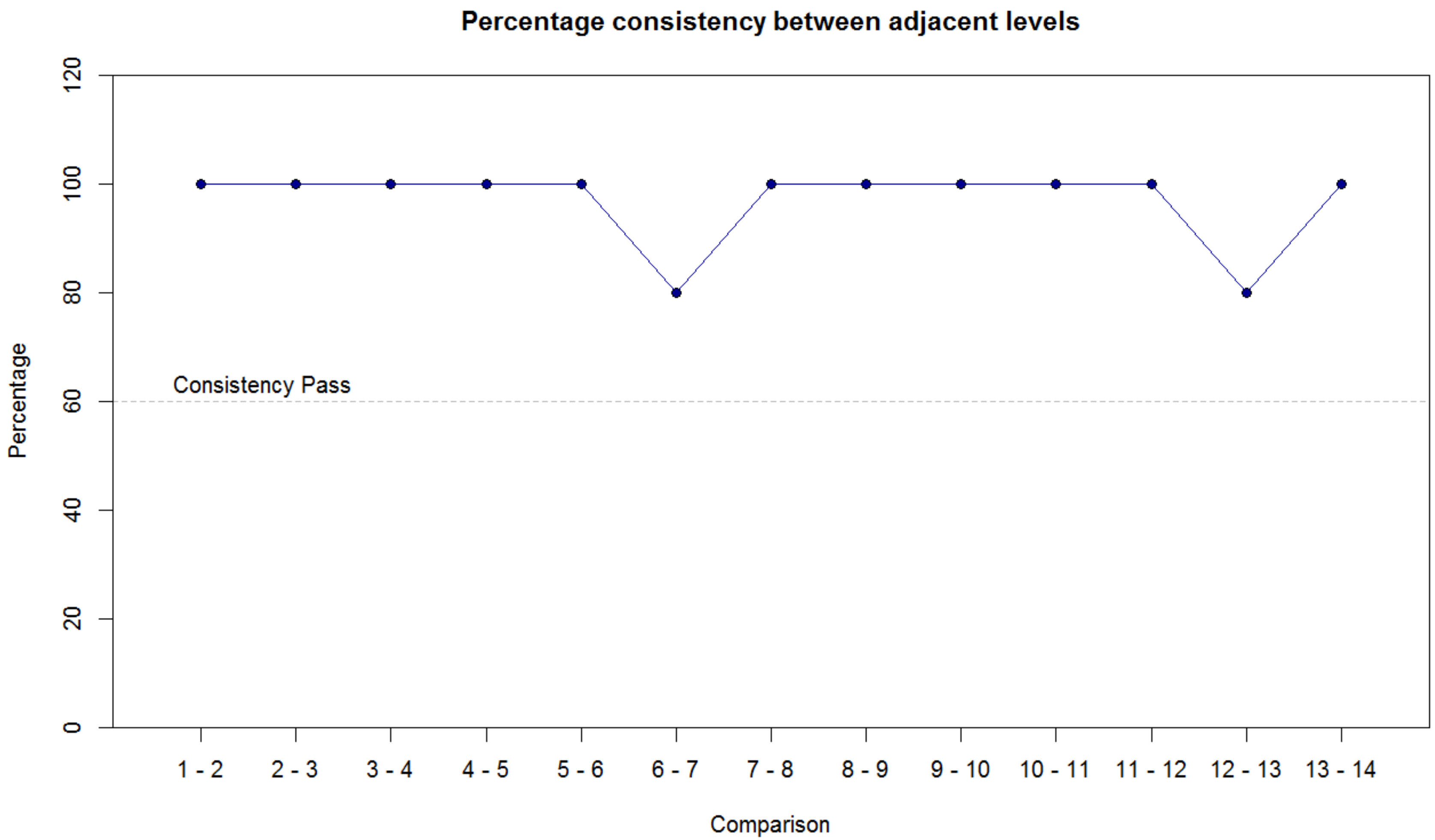
Consistency Plot

MANGOSOLUTIONS
data analysis that delivers



autoPricing Package

Difference between adjacent levels



Summary

Round-up of presentation

MANGOSOLUTIONS
data analysis that delivers

autoPricing allows users to automate GLM pricing by:

- Using a framework built from the forward/backward stepwise algorithms and information criterion selection
- Automating the rebalancing of rating factors by using a combination of mapping tables and pairwise Tukey (HSD) tests
- Create a formal definition of consistency that is numerical and can be included in an automated algorithm
- Incorporating all the above in a decision making algorithm along with standard actuarial model options to arrive at risk and severity model derived the stepwise algorithm
- Allows reports of the fitting process to be generating automatically during the fitting process

Summary

Future work

Develop the package in partnership with users in the insurance market

Actuarial pricing automation working group?

Adding more features:

- Auto-spline fitting
- Pure premium & claims inflation calculations
- Model combiner

Developing a GUI front