Predictive Modeling

And Its Application in Actuarial

Richard Xu, **PhD FSA**

VP & Actuary, Head of Data Science

Global R&D, RGA

June, 2015 **India**

Contents

- Predictive Modeling for Actuary
- Model Specification
- Example – Lapse Model
- Data & Model Considerations

Opportunity

Sales & Marketing

- Customer response modeling ("propensity to buy/renew")
- Recommendations
- Agent recruiting, quality assessment & monitoring

Risk Selection / Risk Scoring

- Predictive Underwriting
- Underwriting triage
- Risk quantifying & risk segmentation
- Improve placement rates

Pricing / Product Development

- More pricing variables & more accurate
- Better incorporated interaction
- Price optimization
- More accurate formula-driven assumptions

Experience Analysis

- Mortality, lapse, incidence, etc
- True multivariate approach
- Efficient use of data
- Handle low-credibility data
- Create assumptions

In-force Policy Management

- Customer retention model "propensity to lapse/persist
- Customer lifetime value

Claims Administration

- Claim risk scoring
- Claims triage
- Fraudulent claims
- Rescinded claims

PM for Actuary

- Why actuaries need to bother with PM?
- PM advantage
  - True multivariate approach
    - Eliminate bias of uni-variate
  - Efficient way to use data
    - Better way for low credible data; improved results vs. traditional
  - Statistical results
    - Not only mean/expected values, but also uncertainty
  - Inclusion of interaction term
    - Two-way or higher order for correction of combination of certain variables

PM for Actuary

- GLM - main focus of PM in insurance industry
- Inclusion of most distributions in insurance
  - ✓ Normal, binomial, Poisson, Gamma, inverse-Gaussian, Tweedie
  - ✓ Extension of Ordinary Least Square (OLS)

    OLS $$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n = \Sigma_i \beta_i X_i$$

    GLM $$g(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n = \Sigma_i \beta_i X_i$$
  - ✓ Easy to understand & communicate
- Powerful and flexible
  - ✓ Weights for data credibility & offset for known effects
  - ✓ Non-linear relationships between variables by link function
  - ✓ Multiplicative model intuitive & consistent with actuarial practice

Generalized Linear Model

Distribution	V(μ)	Link	Sample Application
Normal	1	Identity	(LM) General Application
Poisson	μ	Log	Claim frequency/count, experience
Binomial	$\mu(1-\mu)$	Logistic	Retention, cross-sell, UW, experience
Gamma	μ^2	Log	Claim severity
Compound	$\mu^p, p \in (1,2)$	Log	Claim Cost & Premium
Inverse-Gaussian	μ^3	Log	Claim cost

- "Bread and Butter" for PM in insurance
  - ✓ Great flexibility in variance structure
  - ✓ Baseline and intercept
  - ✓ Relatively easy to understand & explain
  - ✓ Good balance between accuracy & interpretability

Model

Methodology

A properly constructed PM model can result with a multiplicative model

$$Rate = Rate_{base} * \prod_i FS_i * \prod_{i,j} FC_{i,j}$$

Rate: mortality, lapse, incidence, termination/continuance, etc.

FS_i is factor for single variable *i* (main effect)

$FC_{i,j}$ is factor for 2-way interaction term of variable *i* and *j*; (higher order possible)

- Consistent with current actuarial practice
- Implement within existing system without overhaul
- Intuitively easy to understand

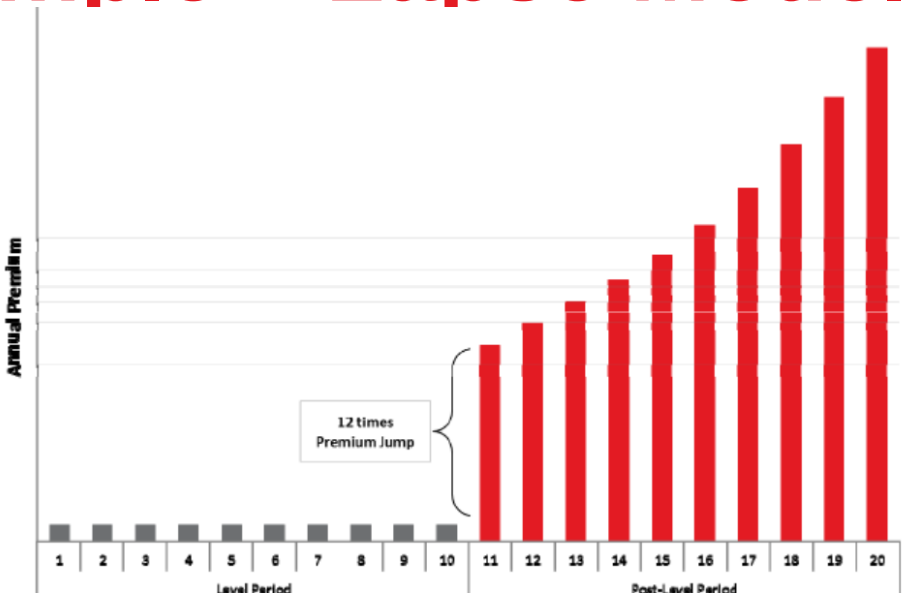
How to build a model

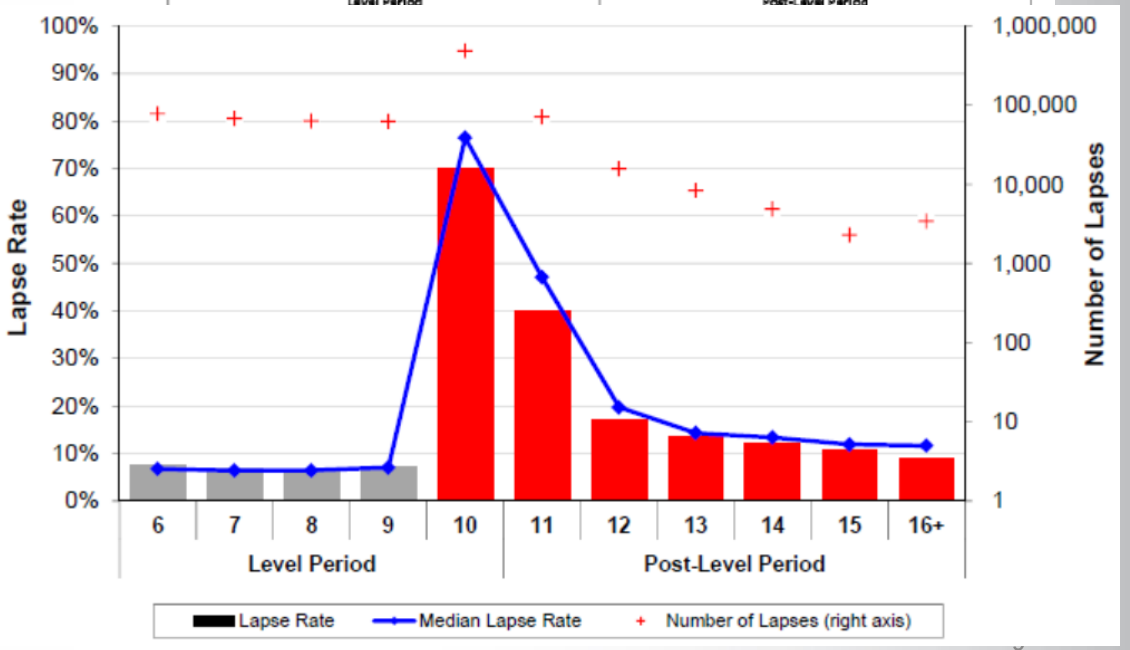
- Set objective
  - ✓ what to achieve by PM
- Process Data
  - ✓ Same as traditional one: understand business & data; clean data; transform data for experience study
  - ✓ May need to split data into development & validation subset
- Fit a model
  - ✓ Select proper distribution for target variable; choose explanatory variables; determine if cross-terms are needed; assess model
  - ✓ Validate the model with validation dataset
- Interpret model & implement
  - ✓ Understand model; extract business insights; implement in business process
- Monitor & Update

Example – Lapse Model

Post-level Term Lapse

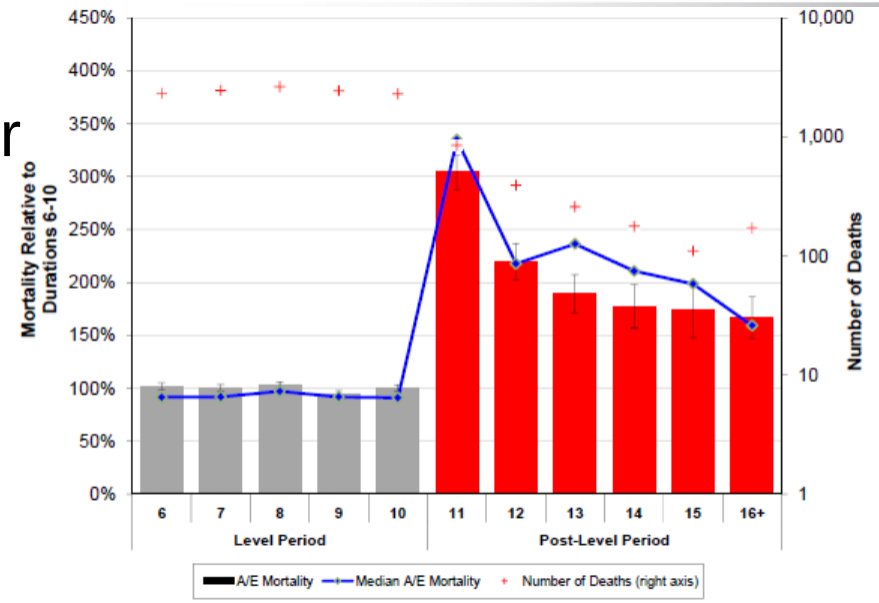
- ✓ SOA research project on PLT lapse study conducted by RGA
- ✓ Account for about 2/3 of US term product sales
- ✓ 10-Year Term: Sample Premium Structure; traditional “Jump to ART”





Example – Lapse Model

- When combined with term tail mortality, it is a powerful tool for optimization
- Lapse model data
  - ✓ Data on T10 for duration 10
  - ✓ Exposure about 690K years & lapse 480K
  - ✓ Variables: age, sex, UW risk, premium mode, premium jump ratio, exposure, lapse, face amount, distribution, base/rider, billing, etc.
  - ✓ Dataset split into two parts, 70% for model development, & 30% for model validation
- Assumptions
  - ✓ You are an actuary, knowing goals, business & data
  - ✓ Data: cleaned, understood & processed

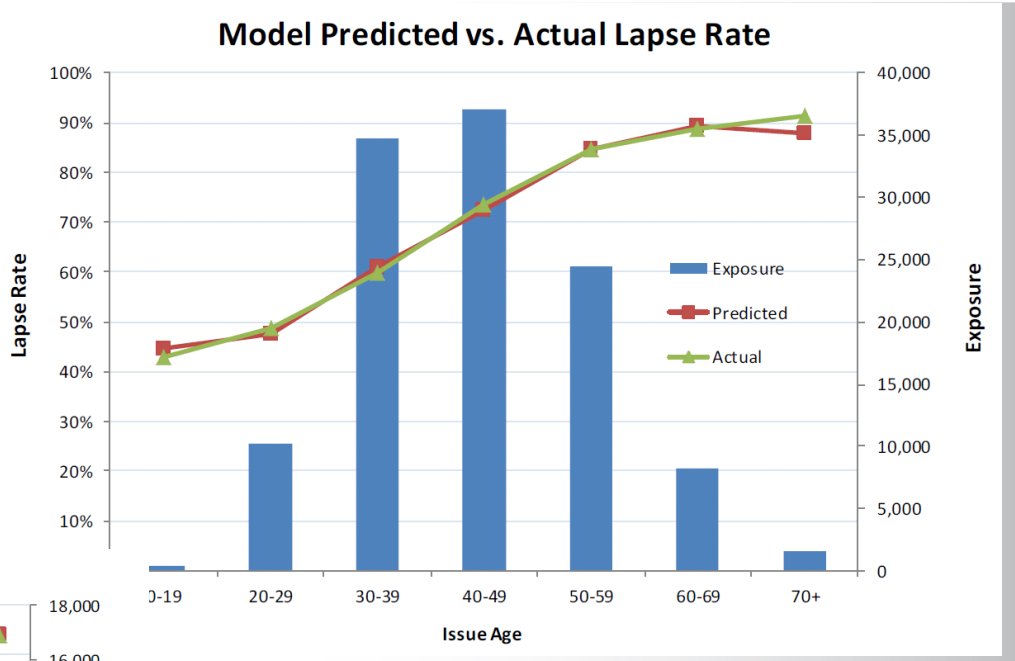


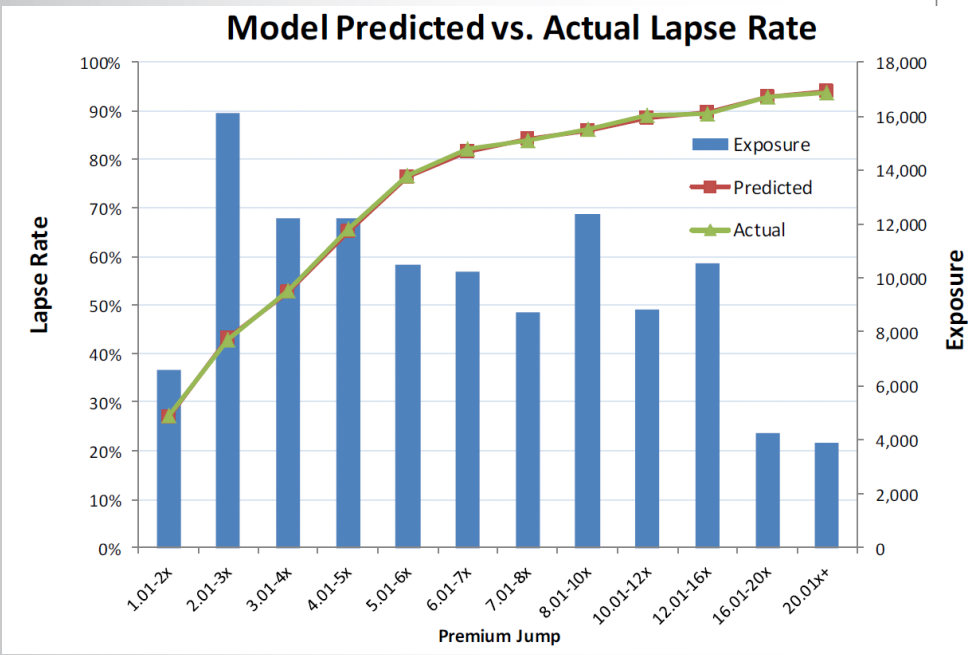
Model Results

Model Parameter						Validation Results			
Variable		Type	Coefficient	P-value	Factor	%	Actual	Predicted	A/E
Intercept		-	3.246	2.03E-14					
Issue Age	Issue Age	Numerical	1.621E-01	<2.00E-16					
	(Issue Age)^2	Numerical	-6.419E-04	<2.00E-16					
	log(Issue Age)	Numerical	-2.725	<2.00E-16					
Risk Class	Super-Pref NS	Categorical	0		**1.00**	17.0%	82.4%	82.5%	100%
	NS		0.03427	1.59E-09	**1.03**	70.5%	68.7%	68.2%	101%
	SM		0.1205	<2.00E-16	**1.13**	12.5%	67.4%	68.3%	99%
Face Amount	<50K	Categorical	0		**1.00**	0.3%	49.5%	55.7%	89%
	50-100K		0.3153	3.49E-15	**1.37**	6.4%	63.4%	63.0%	101%
	100K-250K		0.3437	<2.00E-16	**1.41**	43.9%	69.2%	68.9%	100%
	250K-1M		0.3652	<2.00E-16	**1.44**	41.2%	72.3%	72.1%	100%
	>1M		0.3645	<2.00E-16	**1.44**	8.2%	79.0%	79.3%	100%
Premium Mode	Annual	Categorical	0		**1.00**	22.8%	85.5%	85.0%	101%
	Semi/Quarter		-0.03244	1.16E-11	**0.97**	39.8%	76.1%	75.8%	100%
	Monthly/BiWeekly		-0.2755	<2.00E-16	**0.76**	34.4%	53.3%	53.5%	100%
	Other/Unknown		0.02057	0.0586	**1.02**	3.0%	91.1%	90.5%	101%
Premium Jump	PREM_JUMP 1-2	Categorical	0		**1.00**	5.6%	27.3%	26.9%	102%
	PREM_JUMP 2-3		1.135	<2.00E-16	**3.11**	13.8%	42.8%	43.0%	99%
	PREM_JUMP 3-4		1.492	<2.00E-16	**4.45**	10.5%	52.9%	52.6%	100%
	PREM_JUMP 4-5		1.826	<2.00E-16	**6.21**	10.5%	65.7%	65.1%	101%
	PREM_JUMP 5-6		2.082	<2.00E-16	**8.02**	9.0%	76.7%	76.4%	100%
	PREM_JUMP 6-7		2.118	<2.00E-16	**8.31**	8.8%	82.3%	81.7%	101%
	PREM_JUMP 7-8		2.176	<2.00E-16	**8.81**	7.5%	84.0%	84.1%	100%
	PREM_JUMP8-10		2.246	<2.00E-16	**9.45**	10.6%	86.2%	85.9%	100%
	PREM_JUMP 10-12		2.304	<2.00E-16	**10.01**	7.6%	89.0%	88.5%	101%
	PREM_JUMP 12-16		2.342	<2.00E-16	**10.40**	9.1%	89.4%	89.7%	100%
	PREM_JUMP 16-20		2.385	<2.00E-16	**10.86**	3.6%	92.8%	92.8%	100%
	PREM_JUMP >20		2.356	<2.00E-16	**10.55**	3.4%	93.7%	93.9%	100%
Cross Term	Issue Age & PREM_JUMP	Mixed							

Model Results

- Validation results: model predictive power in real life business
- Lapse rates vs. age and premium jump ratio



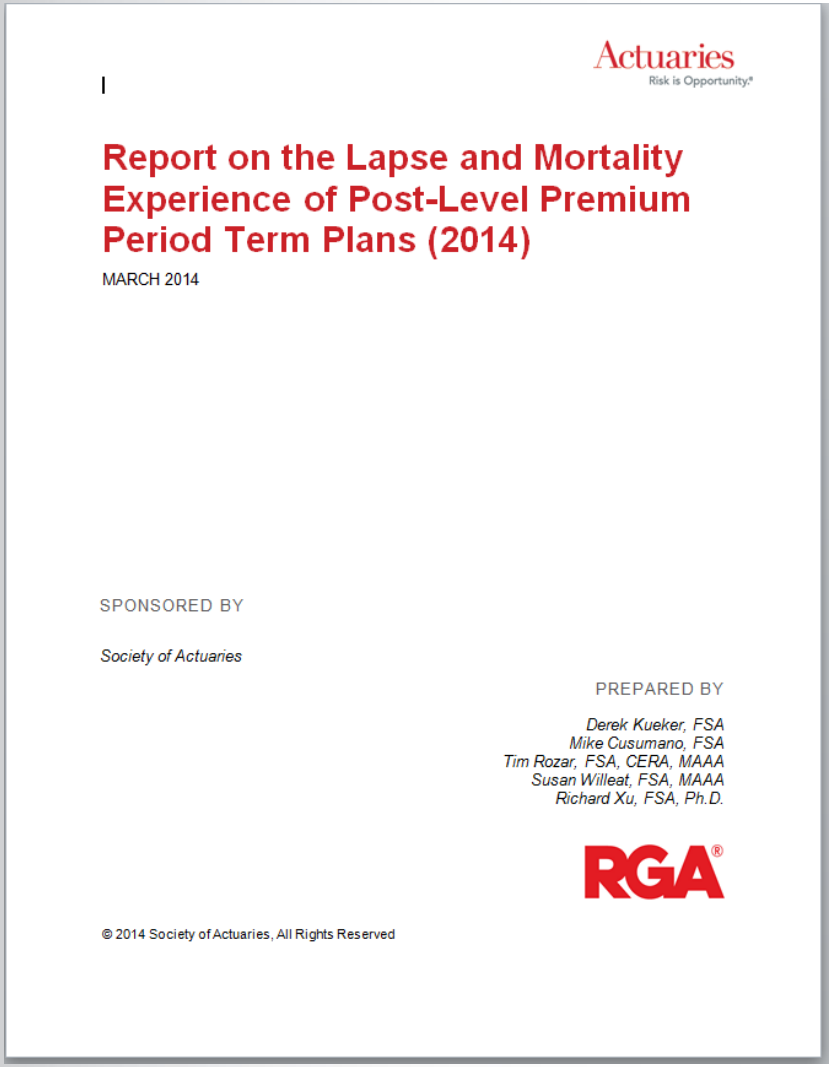


- Data is the key

Example – Lapse Model

```
# 1. load data into R
lapseData <-
read.csv("SampleData2014SOAPM.csv")

# 2. explore data: summary, read first 6 records
summary(lapseData)
head(lapseData)

#other cmnds to explore: data field list, size, tail
names(lapseData)
dim(lapseData)
tail(lapseData)
aggregate(LapsedN ~RiskClass, data=lapseData, sum)

#  3. build a model
Model1 <-
glm(LapsedN~offset(log(Exposure))+FaceAmount+Pr
emiumMode+RiskClass+IssueAge,
family=poisson(),data=lapseData)
summary(Model1)

# include a cross term to improve the model
Model2 <-
glm(LapsedN~offset(log(Exposure))+FaceAmount+Pr
emiumMode+RiskClass+IssueAge+PremiumMode:Iss
ueAge, family=poisson(),data=lapseData)
anova(Model1,Model2)

# 4. predicted values
lapseData$pred <-predict(Model1,lapseData,
type="response")

# 5. prepare data
byPred <- aggregate(pred ~
PremiumMode+RiskClass, data = lapseData, FUN
= sum)
byObsv <- aggregate(LapsedN ~
PremiumMode+RiskClass, data = lapseData, FUN
= sum)
AERatio <- byObsv[,3]/byPred[,3]

#make plot
plot(AERatio,xlab="PremiumMode+RiskClass",
ylab="AE Ratio",xaxt='n',ylim=c(0.9,1.1),pch=18)
title("A/E vs. Premium Mode and Risk Class")
axis(1, at=1:4,labels=c("NS-Annual","NS-
Monthly","SM-Anual","SM-Monthly"), las=0)
abline(1,0,col="red")

# 6. export the data to a csv file
write.csv(lapseData,"modelDataFile.csv")
```

Example



Actuaries
Risk is Opportunity®

Report on the Lapse and Mortality Experience of Post-Level Premium Period Term Plans (2014)

MARCH 2014

SPONSORED BY

Society of Actuaries

PREPARED BY

Derek Kueker, FSA
Mike Cusumano, FSA
Tim Rozar, FSA, CERA, MAAA
Susan Willeat, FSA, MAAA
Richard Xu, FSA, Ph.D.

RGA®

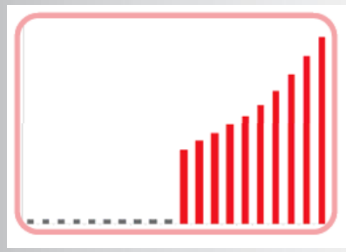
- Final results with detailed model were published by SOA
  - ✓ In SOA website
  - ✓ More focus on traditional analysis with PM as supplement
- PM can add more values than this?

Application

- Policyholder Behavior in the Tail
  - **Causes and Effects**

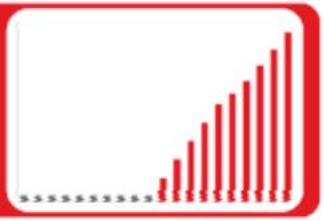
- Higher Premium Jumps
- Increased Lapses
- Greater Anti-selection
- Elevated Mortality & Volatility

- Alternatives to “Cliff” Premium Jump
  - **Term Tail Rescue Options**



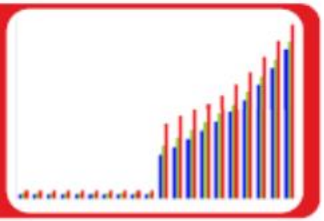
Graded Approach

- Lower initial premium jumps
- Premium grades to ultimate scale over set period



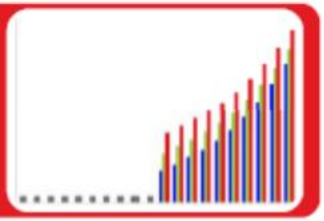
Continued Class Approach

- Class structure continues into post-level period
- Alternative to using residual or undifferentiated rates



Simplified Underwriting Approach

- Simplified UW application near end of level period
- Premium jump based on UW results



Considerations

- Data
  - ✓ #1 issue in PM, just as in traditional study
  - ✓ Understand, clean and process
  - ✓ Missing values; grouping of categories
  - ✓ Too much data, in “big data” territory (nice problem to have)
- Model
  - ✓ A simple model is always better than a complicated one
  - ✓ Over-fitting issue
  - ✓ Validation procedure
- Actuaries are data expert on insurance business
  - Unique position to take up PM for actuarial work
  - Need more training/study on modeling

Predictive Modeling

And Its Application in Actuarial

Richard Xu, **PhD FSA**

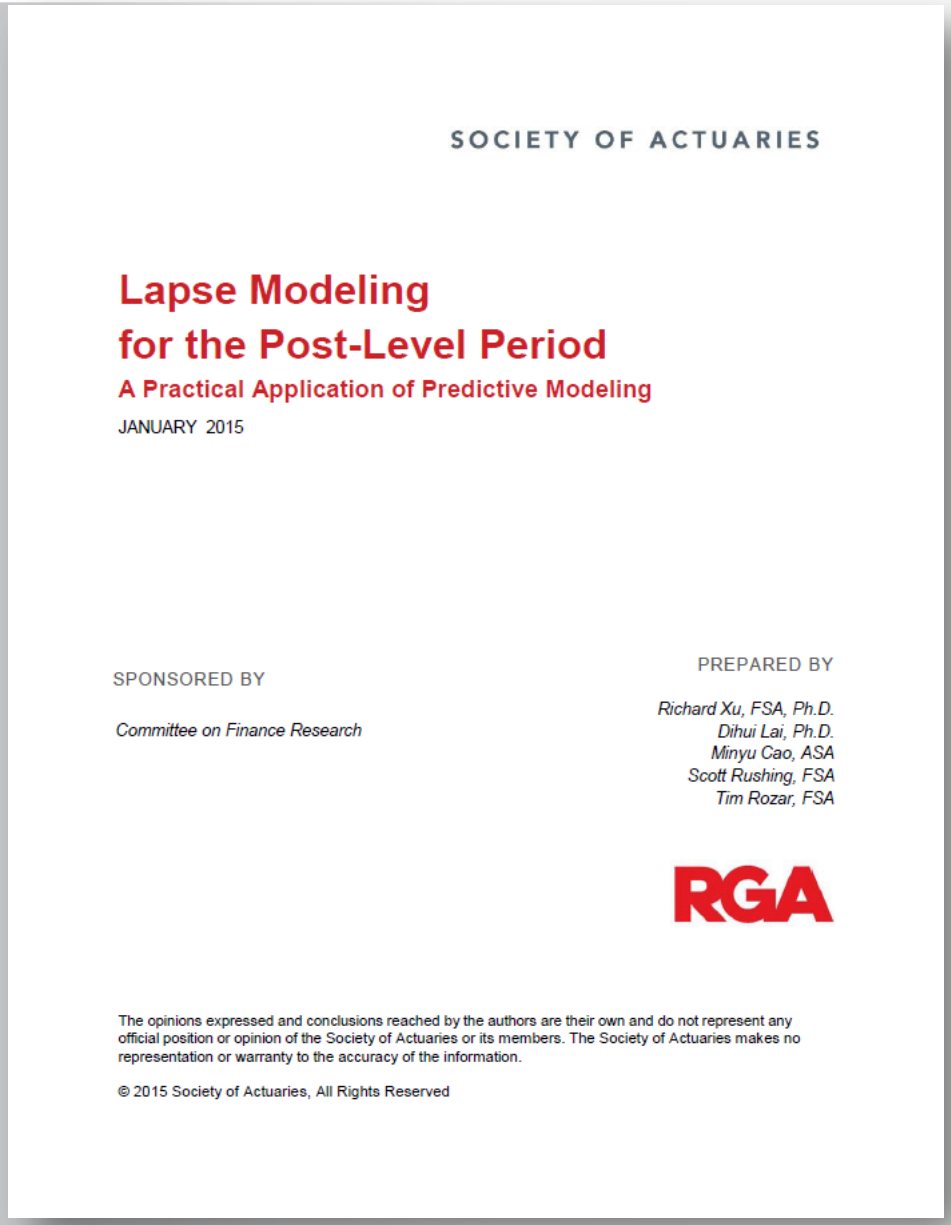
VP & Actuary, Head of Data Science

Global R&D, RGA

June, 2015

India

Lapse Model



SOCIETY OF ACTUARIES

Lapse Modeling
for the Post-Level Period

A Practical Application of Predictive Modeling

JANUARY 2015

SPONSORED BY

Committee on Finance Research

PREPARED BY

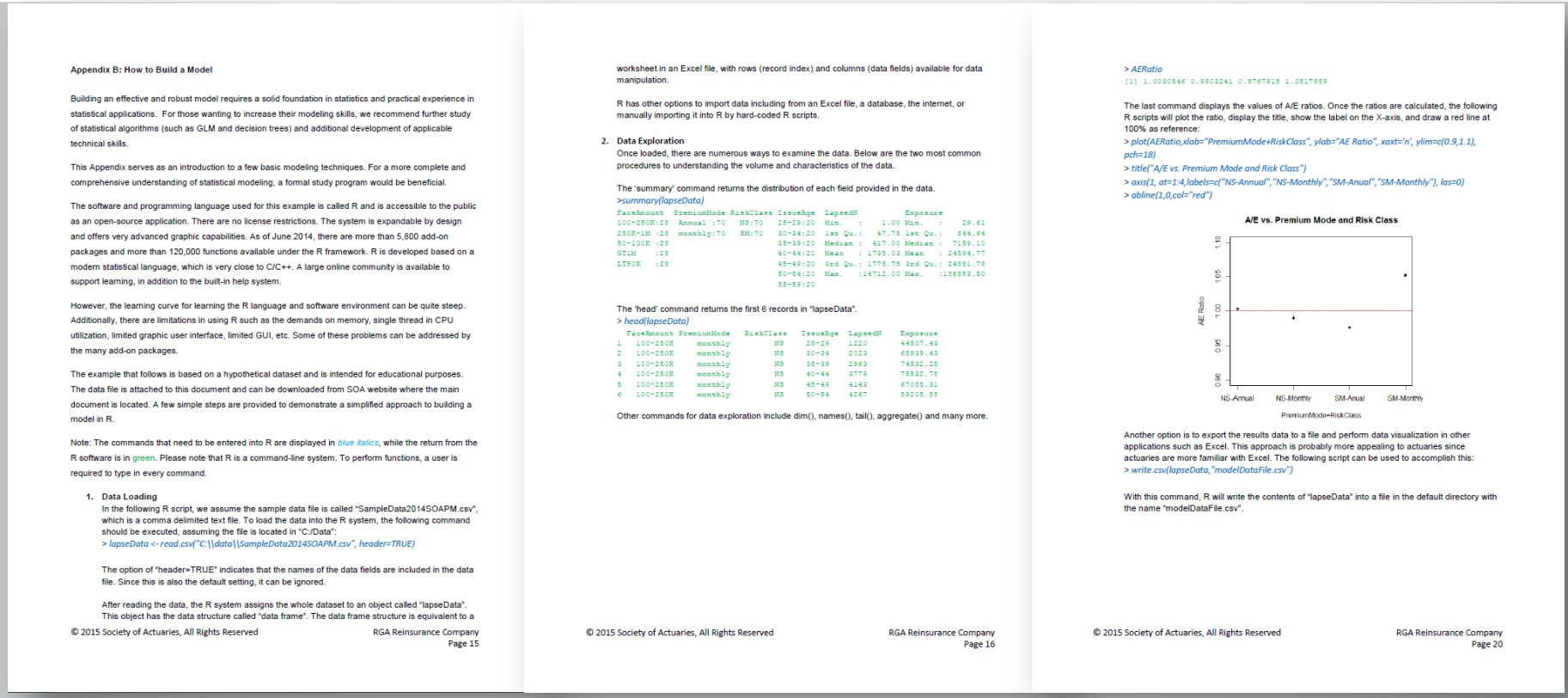
Richard Xu, FSA, Ph.D.
Dihui Lai, Ph.D.
Minyu Cao, ASA
Scott Rushing, FSA
Tim Rozar, FSA

RGA

The opinions expressed and conclusions reached by the authors are their own and do not represent any official position or opinion of the Society of Actuaries or its members. The Society of Actuaries makes no representation or warranty to the accuracy of the information.

- In response to SOA RFP on PM application in insurance, a separate research paper was published
  - ✓ Specifically target PM application
  - ✓ A full lapse model for PLT life product
  - ✓ Not only a industry model/table for PLT lapse rates, but also an educational document

Lapse Model



Appendix B: How to Build a Model

Building an effective and robust model requires a solid foundation in statistics and practical experience in statistical applications. For those wanting to increase their modeling skills, we recommend further study of statistical algorithms (such as GLM and decision trees) and additional development of applicable technical skills.

This Appendix serves as an introduction to a few basic modeling techniques. For a more complete and comprehensive understanding of statistical modeling, a formal study program would be beneficial.

The software and programming language used for this example is called R and is accessible to the public as an open-source application. There are no license restrictions. The system is expandable by design and offers very advanced graphic capabilities. As of June 2014, there are more than 5,800 add-on packages and more than 120,000 functions available under the R framework. R is developed based on a modern statistical language, which is very close to C/C++. A large online community is available to support learning, in addition to the built-in help system.

However, the learning curve for learning the R language and software environment can be quite steep. Additionally, there are limitations in using R such as the demands on memory, single thread in CPU utilization, limited graphic user interface, limited GUI, etc. Some of these problems can be addressed by the many add-on packages.

The example that follows is based on a hypothetical dataset and is intended for educational purposes. The data file is attached to this document and can be downloaded from SOA website where the main document is located. A few simple steps are provided to demonstrate a simplified approach to building a model in R.

Note: The commands that need to be entered into R are displayed in *blue italics*, while the return from the R software is in green. Please note that R is a command-line system. To perform functions, a user is required to type in every command.

1. **Data Loading**
   In the following R script, we assume the sample data file is called "SampleData2014SOAPM.csv", which is a comma delimited text file. To load the data into the R system, the following command should be executed, assuming the file is located in "C:/Data":
   > lapseData <- read.csv("C:\\data\\SampleData2014SOAPM.csv", header=TRUE)

   The option of "header=TRUE" indicates that the names of the data fields are included in the data file. Since this is also the default setting, it can be ignored.

   After reading the data, the R system assigns the whole dataset to an object called "lapseData". This object has the data structure called "data frame". The data frame structure is equivalent to a

© 2015 Society of Actuaries, All Rights Reserved RGA Reinsurance Company Page 15

worksheet in an Excel file, with rows (record index) and columns (data fields) available for data manipulation.

R has other options to import data including from an Excel file, a database, the internet, or manually importing it into R by hard-coded R scripts.

2. **Data Exploration**
   Once loaded, there are numerous ways to examine the data. Below are the two most common procedures to understanding the volume and characteristics of the data.

   The 'summary' command returns the distribution of each field provided in the data.
   >summary(lapseData)

```
FaceAmount  PremiumMode RiskClass IssueAge  LapsedN          Exposure
100-250K:28  Annual :70   NS:70   25-29:20  Min.   :   1.00 Min.   :    29.61
250K-1M :28  monthly:70   SM:70   30-34:20  1st Qu.:  47.75 1st Qu.:   644.64
50-100K :28                       35-39:20  Median :  417.00 Median :  7159.10
GT1M    :28                       40-44:20  Mean   : 1738.03 Mean   : 24594.77
LT50K   :28                       45-49:20  3rd Qu.: 1775.75 3rd Qu.: 24881.78
                                  50-54:20  Max.   :14712.00 Max.   :186883.80
                                  55-59:20
```

   The 'head' command returns the first 6 records in "lapseData".
   > head(lapseData)

```
  FaceAmount PremiumMode   RiskClass   IssueAge  LapsedN    Exposure
1   100-250K     monthly          NS     25-29     1220    44507.43
2   100-250K     monthly          NS     30-34     2023    65929.43
3   100-250K     monthly          NS     35-39     2963    74532.25
4   100-250K     monthly          NS     40-44     3779    75532.78
5   100-250K     monthly          NS     45-49     4143    67085.31
6   100-250K     monthly          NS     50-54     4267    59205.88
```

   Other commands for data exploration include dim(), names(), tail(), aggregate() and many more.

© 2015 Society of Actuaries, All Rights Reserved RGA Reinsurance Company Page 16

> AERatio

```
[1] 1.0030546 0.9903241 0.9767918 1.0517889
```

The last command displays the values of A/E ratios. Once the ratios are calculated, the following R scripts will plot the ratio, display the title, show the label on the X-axis, and draw a red line at 100% as reference:

> plot(AERatio,xlab="PremiumMode+RiskClass", ylab="AE Ratio", xaxt='n', ylim=c(0.9,1.1), pch=18)
> title("A/E vs. Premium Mode and Risk Class")
> axis(1, at=1:4,labels=c("NS-Annual","NS-Monthly","SM-Anual","SM-Monthly"), las=0)
> abline(1,0,col="red")

Another option is to export the results data to a file and perform data visualization in other applications such as Excel. This approach is probably more appealing to actuaries since actuaries are more familiar with Excel. The following script can be used to accomplish this:
> write.csv(lapseData,"modelDataFile.csv")

With this command, R will write the contents of "lapseData" into a file in the default directory with the name "modelDataFile.csv".

© 2015 Society of Actuaries, All Rights Reserved RGA Reinsurance Company Page 20

- Appendix of the report include an example
  - ✓ Step-by-step on how to build a real lapse model in R
  - ✓ A sample data file and R script file are include
  - ✓ All can be downloaded from SOA website