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.

TABLE OF CONTENTS

Introduction

The Society of Actuaries (SOA) selected RGA Reinsurance Company (RGA) to undertake a research project to demonstrate an application of predictive modeling to life insurance. The goal was to illustrate an application of predictive modeling (PM) to the post-level period for 10-year level term insurance and model lapses in a multivariate setting. The model presented in this paper is an extension of previous work completed for the May 2014 “Report on the Lapse and Mortality Experience of Post-Level Premium Period Term Plans,” which is available on the SOA website (https://www.soa.org/Research/Experience-Study/Ind-Life/Persistency/research-2014-post-level-shock.aspx). The May 2014 study is also a follow-up to SOA-sponsored research completed by RGA in July 2010 “Lapse and Mortality Experience of Post-Level Premium Period Term Plans”.

The original model presented in the post-level term experience report was designed to be an introductory predictive model and focused only on duration 10 shock lapse for 10-year term products. The model incorporated a variety of correlated predictor variables including age, premium jump, premium payment mode and face amount.

The model presented in this report is an improvement over the previous model as it expands our look at and use of potential predictor variables. Variable transformations, higher order terms and cross-terms have also been introduced into the model improving the overall model fit. This model also extends beyond the initial shock lapse all the way to duration 19. This paper additionally provides a more comprehensive discussion of the model development process, as well as sample code to provide further education to the reader.

Special Thanks

The authors would like to thank the SOA and the members of the Project Oversight Group (POG) for their guidance and support on this research project. The POG consists of the following members: William Cember, Andy Ferris, Jean-Marc Fix (chair), John Hegstrom, Christine Hofbeck, Steve Marco, Dennis Radliff, Barbara Scott (SOA), Steven Siegel (SOA). Their comments, feedback and direction have greatly improved the overall value of this project.

In addition, the authors express our sincere thanks to Mike Cusumano, Derek Kueker and Kent Wu of RGA for their valuable feedback and contributions to the final model.

Executive Summary

Advanced analytical tools such as predictive modeling can deliver improved insight into the understanding of a wide array of real-world problems. As it relates specifically to this paper, the use of predictive modeling provides a richer context for exploring the interaction between product design and policyholder behavior. As described in detail in the 2014 "Report on the Lapse and Mortality Experience of Post-Level Premium Period Term Plans", level-term products in the U.S. typically include a very large increase in the premium rate as the policy transitions from the level period to the post-level period. This jump in premium creates a decision for policyholders. They can either pay the much higher premiums to continue coverage or they can lapse their policies and perhaps seek more affordable newly-underwritten coverage. As expected, a large proportion of policyholders choose the latter option and this product is characterized by high shock lapses at the end of the level period. Correspondingly, the mortality levels tend to be much higher in the post-level period as those who could no longer qualify for a new policy are disproportionately likely to pay the higher premium rates.

This paper takes an in-depth look at how generalized linear regression can be used to predict the annual lapse rates on term policies beyond the level premium period. The models developed demonstrate not only the predictive value of various policy attributes on lapse rate, but also the nuanced interaction between these variables. The relationship between issue age, premium jump and policy duration are investigated in detail as these are the three most significant independent predictors of post-level period lapsation. Face amount, premium payment mode and risk class are also demonstrated to add lift to the final model.

Another key objective of this paper is to provide educational background on the process of building predictive models. The following key topics are described herein:

- Data Preparation and Quality: This is the most important part of any model building process. The models built for this paper are supported heavily by the in-depth data analysis, data scrubbing and business knowledge provided by the 2014 SOA post-level term research team.
- Model selection and validation: A variety of model forms and predictor variables are often available to the data scientist when building a model. This paper describes the use of iterative variable selection using the Akaike information criterion (AIC). In addition, the use of variable transformations and higher power terms are introduced to further improve model performance.
- Model interpretation: This paper provides definitions of many terms that are used to interpret a model's performance. Visual and tabular displays are also provided to compare model predictions with out-of-sample validation results.
- Model implementation: A sample spreadsheet calculation is provided to illustrate how the results from a model can be used directly in the assumption development process for pricing or reserving.

- Sample coding in R: At the request of the Project Oversight Group, this paper provides specific guidance for building models using R including sample coding that can be used to load and profile data, build and validate models, and display results visually.

Data

To support the creation of the model discussed in this paper, we employed the same 37 company industry dataset used to create the 2014 SOA / RGA post-level term experience study. The exposures and lapses from that 2000-2012 industry study were the primary sources of data used for the creation of the models provided in this paper. The scope of the model presented in this paper was restricted to the post-level term business for 10-year level term plans with a “jump to ART” post-level premium structure. For a more in depth understanding of the source data and post-level term lapses and mortality, please refer to RGA’s paper “Report on the Lapse and Mortality Experience of Post-Level Premium Period Term Plans” on the SOA website.

Data Processing

Due to the rigorous work of cleaning and preparing the data for the original post-level term mortality and lapse studies, the data used in this project was essentially ready for the creation of the predictive models. More about the original data preparation process can be found in the paper provided on the SOA website. Only minor adjustments were made to the original data before the post-level term models were built.

There are two types of variables in modeling, categorical and numerical. Variables such as issue age and duration are defined as numerical variables. Some variables such as risk class and premium mode are categorical in nature, while other variables such as face amount are due to data being previously grouped. Categorical variables can be converted into numerical variables when needed. For example, the premium jump ratio was originally provided as a categorical variable. For the model, we converted premium jump back into numeric values by using the center of each band in order to treat it as a continuous variable in the model. This assumes that lapse rates will continuously increase as the size of the premium jump increases. Moreover, this approach will reduce the number of variables in the final model and effectively reduce the possibility of overfitting the model to the data. The term “overfitting” is used to describe models that are too complex relative to the amount of modeling data available. Models that are overfit to the data actually lose predictive power. Overfitting is one of the primary challenges data scientists face when building effective models.

Other categorical variables were regrouped in many cases in order to get an amount of exposure for each sub-category that can produce statistically significant results (e.g. semiannual and quarterly payments are grouped together for the predictor variable ‘Premium Mode’).

Modeling Approach

A Generalized Linear Model (GLM) assumes the response variable in a dataset follows a distribution in the exponential family (including normal, Poisson, gamma distribution etc.). This approach allows the estimation of an otherwise nonlinear system to occur in a linear fashion. In order to model and understand the lapse behavior of the policyholders under consideration, we assume the number of lapses follows a Poisson distribution. The default link function for Poisson distribution is logarithmic, which could produce a multiplicative model to calculate estimated lapse rate. The multiplicative model is consistent with current actuarial practice and produces a somewhat intuitive result.

Under a GLM framework, the expected occurrence of lapse can be formulated as: $\mu_i = \exp(x_i \beta)$ where

μ_i denotes the estimated lapse rate for the i^th record;

$x_i = (x_{i0}, x_{i1}, x_{i2} ... x_{ip})$ are contributing predictors such as issue age, duration, face amount etc.;

β is the parameter vector and is optimized by maximizing the log-likelihood function.

Variable Selection Process

Model variables were selected from the initial dataset based on their ability to predict lapse rates. Contributions by specific variables to the overall quality of the model are identified through use of the Akaike information criterion (AIC), which forces a balance between model simplicity and likelihood maximization. A model with smaller AIC usually indicates a better fit of the model. For example, the current final model has an AIC of 862,041. If we include the variable of ‘base/rider indicator’, the AIC increases to 862,044. In this case, the addition of the variable does not improve the model. Several variables such as ‘distribution system’, ‘billing type, and ‘underwriting requirement’ were eliminated from the model due to their minimal impact on the AIC.

Statistics-based criterion alone cannot guarantee an effective and robust predictive model. Accuracy of the data and consistency of the variables over time are also considered when selecting the appropriate variables. Additionally, business knowledge and experience play equally important roles in determining the selection of variables for the final model. For example, the variable ‘calendar year’ initially suggested a significant impact on the lapse rate; however, the calendar year effect is truly caused by the changes in the CSO mortality tables during the last decade. Due to the lack of meaningful predictive power, calendar year was dropped from the model. Likewise, the variable ‘issue year’ was dropped due to a strong correlation with the increases in premium jump seen during the last decade.

Certain transformations of variables are also considered in an effort to enhance model performance. For example, the lapse rates as a function of premium jump have very complicated shapes which cannot be explained by a simple linear predictor. The inverse of premium jump and its higher orders help to describe the reduced effect on the lapse rate when premium jumps are large. Since lapse rates are not a simple exponential function of duration, higher order terms of duration are used to describe the trend of lapse rate as duration increases. Second order cross terms are also investigated to consider the multiplicative nature between the linear predictors, such as issue age and duration.

Model Results

The final variables selected for the model and their corresponding variable type and model coefficients are presented on the left side of Figure 1. The right side of Figure 1 shows the proportion of data for each category within a group as well as the actual lapse rates observed, predicted lapse rates and the actual / predicted ratios. More details of the contents of Figure 1 are provided by the following:

Model Parameter Section:

When developing the model, modeling data is randomly ordered and typically split into both training and validation data. A common approach is to use 60% of the data for training the model and 40% of the data for validating the effectiveness of the model. The model parameter section is on the left side of Figure 1 and focuses on the model created using the training data set.

Coefficient – This is the key parameter produced by the model and is used in the formula to calculate the predicted lapse rates based on policy characteristics for the model variables.

Factor – For the three categorical variables in the model, the factors provided show the impact of one specific category relative to the baseline (indicated by a factor of 1.00). For example, NS is set as the baseline for the categorical variable “risk class”. The factor of 1.11 for SM indicates that smokers generally have lapse rates 11% higher than non-smokers’, given everything else is equal.

P-Value – It is a probability value used in statistical significance testing to help determine the significance of including a specific variable in the model. The closer the P-value is to 0, the more likely the variable belongs in the model.

Variable Types:

- Intercept – This is the constant term in the predicted lapse rate formula. For the Poisson distribution, the model predicted lapse rate = EXP(Coefficient * X + Intercept). For more information, please refer to Appendix A: Sample Policy Lapse Rate Calculation.

- Categorical Variables – These are variables that are not numeric in nature.
- Numerical Variables – These are variables that take on numeric values.
- Cross Terms – The goal is to capture the significant interactive effects that may exist between predictor variables.

Validation Results Section:

The validation section is provided on the right side of Figure 1 and gives the result of applying the final model parameters to the 40% validation portion of the data.

Data Proportion – This shows the proportion of data in a given category by exposure count.

Actual Lapse Rate – This column shows the lapse rate calculated using a traditional experience study approach.

Predicted Lapse Rate – This column shows the lapse rate predicted by the model.

Actual / Predicted – This ratio shows deviation of predicted lapse rates from actual lapse rates.

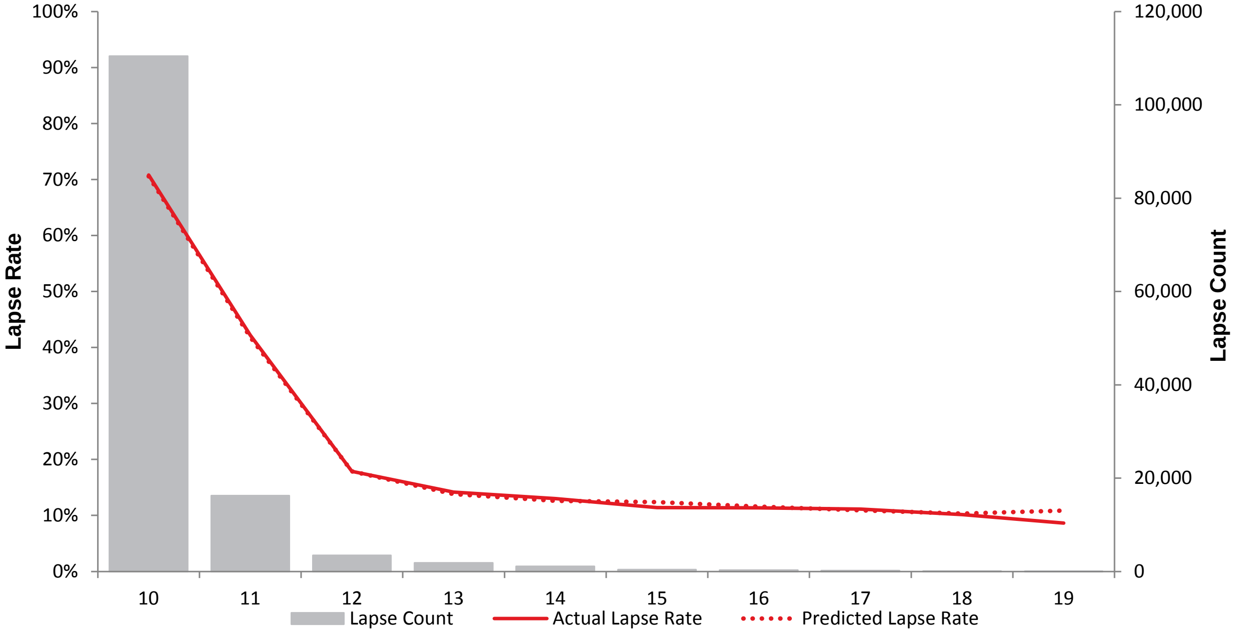
Figure 1: 10 Year Level Term Model Results

	Model Parameter				Validation Results			
Type	Variable	Coefficient	Factor	P-value	Data Proportion (Expos Cnt)	Actual Lapse Rate	Predicted Lapse Rate	Actual/ Predicted
Intercept		5.8348		<2.22E-16				
Categorical								
	Risk Class							
	NS (Other Non-Smoker)	0	1.00	-	74.28%	51.5%	51.4%	100.3%
	BCNS (Best Class Non-Smoker)	-0.0374	0.96	2.42E-08	13.16%	75.0%	74.7%	100.4%
	SM (Smoker)	0.1002	1.11	<2.22E-16	12.57%	53.6%	53.4%	100.4%
	Face Amount							
	<50K	0	1.00	-	0.36%	38.1%	37.2%	102.4%
	50-100K	0.3674	1.44	<2.22E-16	7.61%	43.9%	43.9%	99.9%
	100K-250K	0.4424	1.56	<2.22E-16	45.37%	52.1%	52.1%	100.1%
	250K-1M	0.4827	1.62	<2.22E-16	40.00%	57.8%	57.5%	100.6%
	>1M	0.4906	1.63	<2.22E-16	6.66%	69.0%	68.9%	100.2%
	Premium Mode							
	Monthly	0	1.00	-	41.70%	40.3%	39.7%	101.7%
	Semiannual/Quarterly	0.2221	1.25	<2.22E-16	37.13%	61.5%	61.7%	99.7%
	Annual	0.2264	1.25	<2.22E-16	19.06%	70.2%	70.3%	99.8%
	Other/Unknown	0.2612	1.30	<2.22E-16	2.11%	87.0%	86.9%	100.2%
Numerical								
	Issue Age	0.1270		<2.22E-16				
	(Issue Age)^2	-0.0007		<2.22E-16				
	log(Issue Age)	-2.6857		<2.22E-16				
	(Duration - 9)^(-1)	-12.1912		<2.22E-16				
	(Duration - 9)^(-2)	32.1786		<2.22E-16				
	(Duration - 9)^(-3)	-20.4880		<2.22E-16				
	(Premium Jump Ratio)^(-1)	-2.8684		<2.22E-16				
	(Premium Jump Ratio)^(-2)	-2.9429		<2.22E-16				
	(Premium Jump Ratio)^(-3)	4.0217		<2.22E-16				
Cross Term								
	Issue Age:(Premium Jump Ratio)^(-1)	0.0372		<2.22E-16				
	Issue Age:(Duration - 9)	-0.0032		<2.22E-16				

Results by Duration

Figure 2 demonstrates that the predicted lapse rates from the model very closely follow the actual lapse rates, especially in durations close to the end of the level period. As the number of lapses decrease in later durations, the predicted lapse rates diverge slightly from the actual lapse rates.

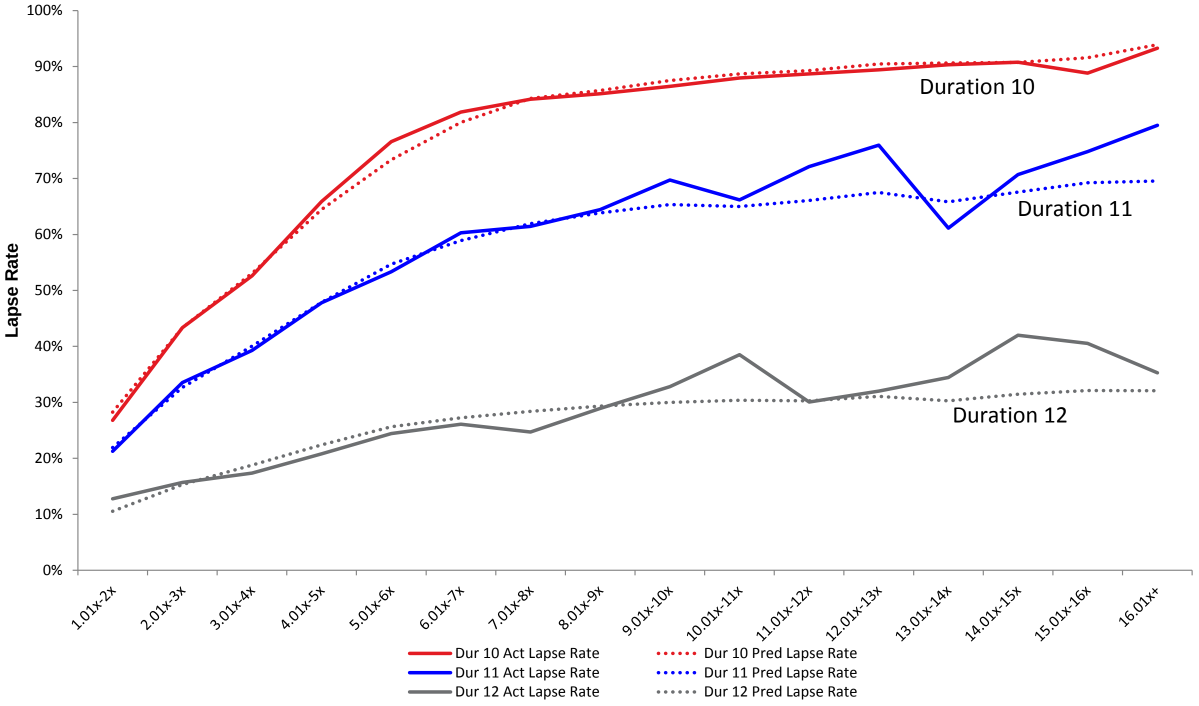
Figure 2: Model Results by Duration



Results by Premium Jump

Figure 3 compares actual and predicted lapse rates for durations 10, 11 and 12. The predicted lapse rates show a smoother trend, whereas the actual lapse rates fluctuate in the areas with the smallest exposure.

Figure 3: Model Results by Premium Jump

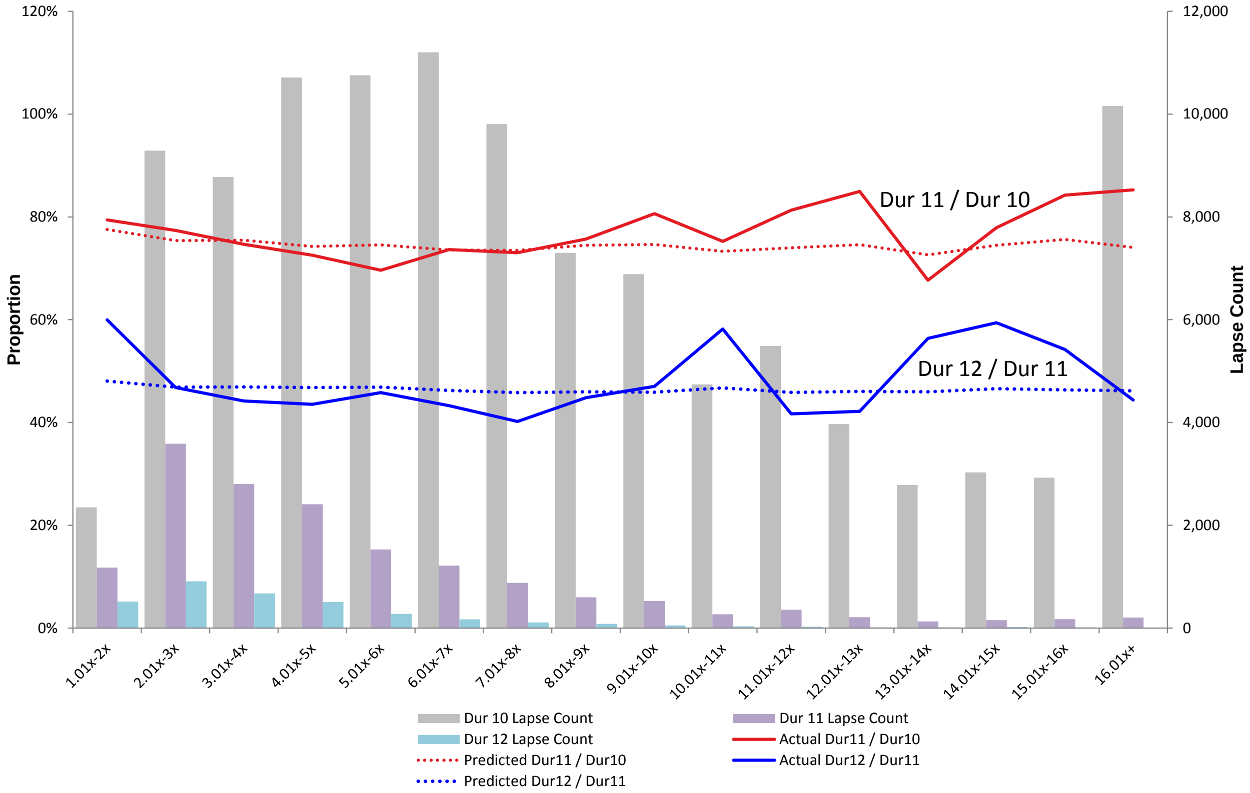


Results by Premium Jump Ratio

Figure 4 compares the ratio of duration 11 lapses to duration 10 lapses, as well as the ratio of duration 12 lapses to duration 11 lapses. The predicted lapse trends shown are nearly level across all premium jump ratios whereas the actual lapse trends are quite volatile where the exposure is thin.

A graph similar to Figure 4 was presented in the May 2014 paper (on page 26) illustrating the relationship below.

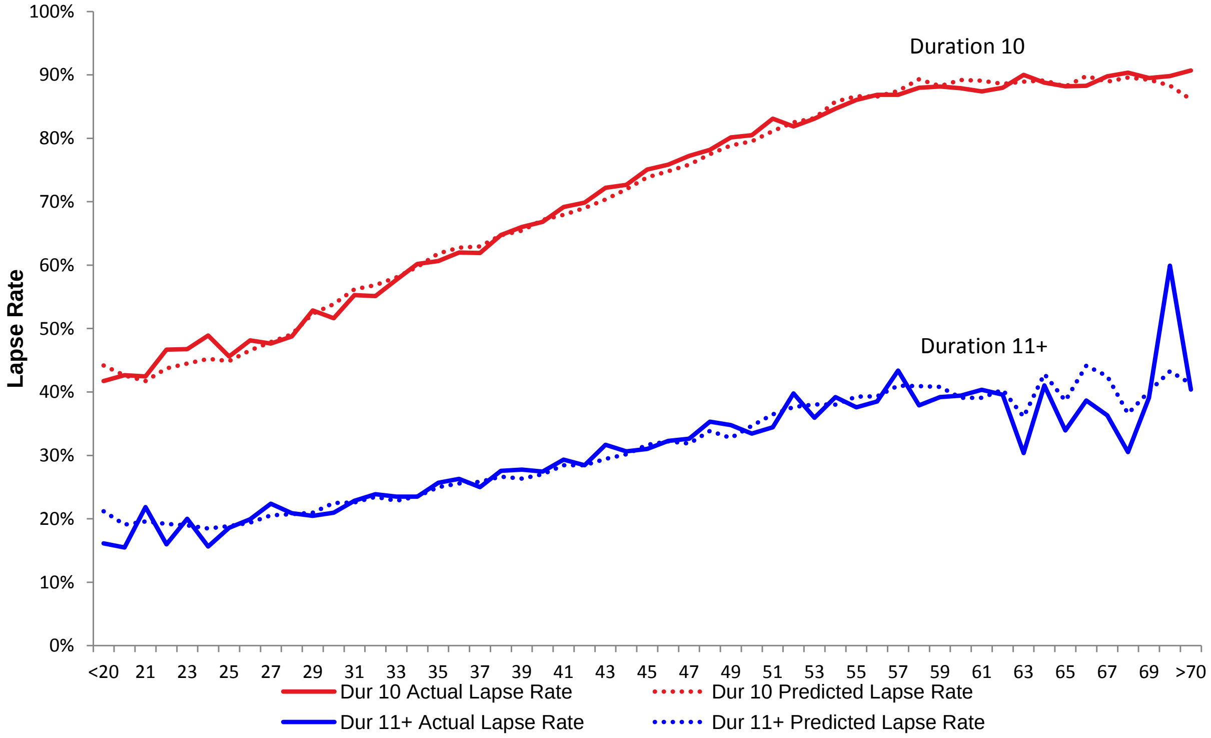
Figure 4: Model Results by Premium Jump Ratio



Results by Issue Age

Figure 5 demonstrates that the predicted lapse rates by issue age are very close to the actual lapse rates. The slight bumpiness in the predicted lapse rates is mostly due to a combination of factors affecting the change in business mix by issue age.

Figure 5: Model Results by Issue Age



Results by Face Amount Band and Duration

Looking at Figure 6a and 6b, the model performs better for higher face amount bands and in areas with more exposure. The relatively weak performance in duration 11 (Figure 6a) and face amount band 50-100k (Figure 6b) are driven primarily by factors outside the model. Although the slight discrepancies could possibly be fixed by introducing additional variables to the model, the modeling team chose model simplicity due to concerns of overfitting the model.

Figure 6a: Model Results by Face Amount Band and Duration

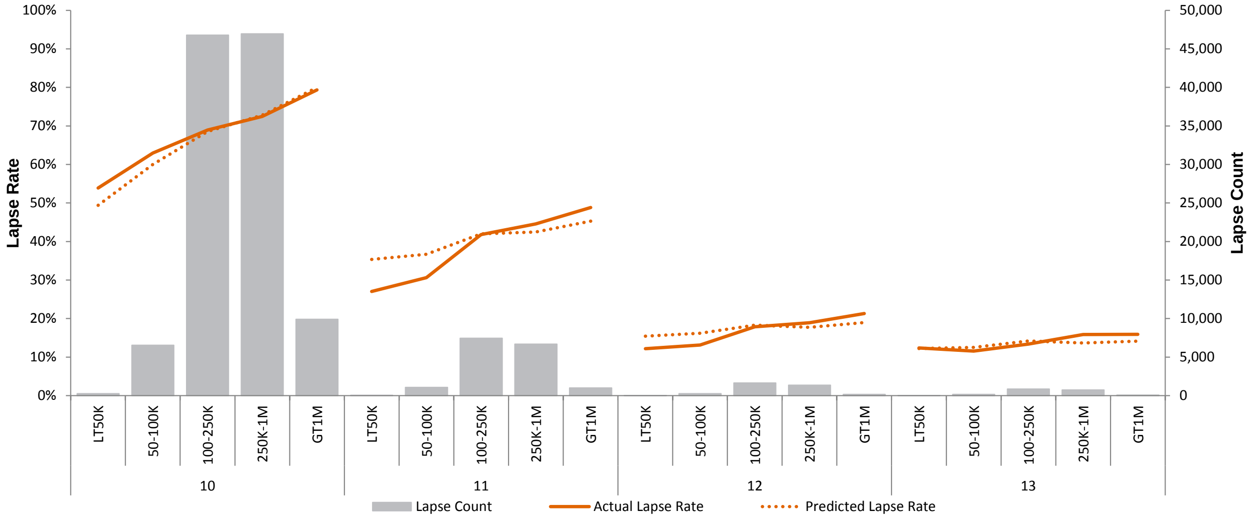
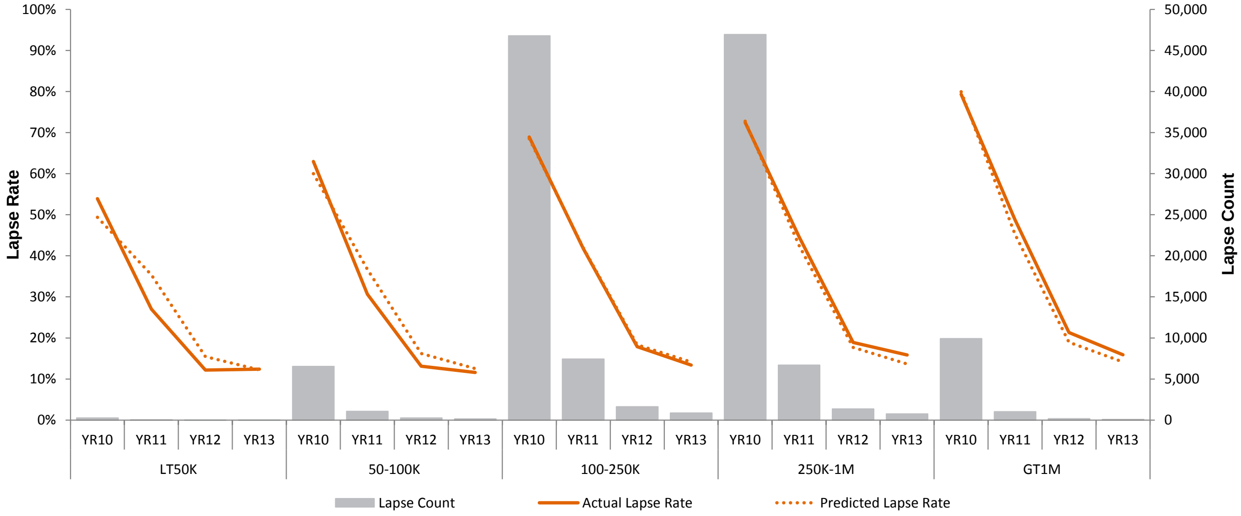


Figure 6b: Model Results by Duration and Face Amount Band



Appendix A: Sample Policy Lapse Rate Calculation

Below is a sample illustration of how to use the model to calculate an expected lapse rate given an individual's issue age, policy duration, risk class, face amount, premium mode, and premium jump. For the particular cell illustrated, the actual lapse rate comes in 2.6% higher than the predicted lapse rate generated by the model. A stand-alone Excel spreadsheet is available on SOA's website for reference to recreate the calculations.

Assumptions	
Issue Age	40
Duration(>=10)	11
Risk Class	NS (Other Non-Smoker)
Face Amount	250K-1M
Premium Mode	Monthly
Premium Jump Ratio	3.01x-4x

Model Variables		Coefficients (a)	Sample Value of x_i (b)	Sample Calculation (c) = (a) * (b)
Intercept		5.8348	1	5.8348
Issue Age		0.1270	40	5.0795
(Issue Age)^2		-0.0007	40^2	(1.0478)
log(Issue Age)		-2.6857	ln(40)	(9.9073)
(Duration - 9)^(-1)		-12.1912	(11-9)^(-1)	(6.0956)
(Duration - 9)^(-2)		32.1786	(11-9)^(-2)	8.0447
(Duration - 9)^(-3)		-20.4880	(11-9)^(-3)	(2.5610)
(Premium Jump Ratio)^(-1)		-2.8684	3.5^(-1)	(0.8195)
(Premium Jump Ratio)^(-2)		-2.9429	3.5^(-2)	(0.2402)
(Premium Jump Ratio)^(-3)		4.0217	3.5^(-3)	0.0938
Risk Class				
	NS (Other Non-Smoker)	0.0000	1	-
	BCNS (Best Class Non-Smoker)	-0.0374	0	-
	SM (Smoker)	0.1002	0	-
Face Amount				
	<50K	0.0000	0	-
	50-100K	0.3674	0	-
	100K-250K	0.4424	0	-
	250K-1M	0.4827	1	0.4827
	>1M	0.4906	0	-
Premium Mode				
	Monthly	0.0000	1	-
	Semiannual/Quarterly	0.2221	0	-
	Annual	0.2264	0	-
	Other/Unknown	0.2612	0	-
Cross Term				
	Issue Age:(Premium Jump Ratio)^(-1)	0.0372	40*3.5^(-1)	0.4247
	Issue Age:(Duration - 9)	-0.0032	40*(11 - 9)	(0.2530)

Results	
Linear Predictor = Sum(Beta$_i$ * x$_i$) = Sum (c)	(0.9642)
Modeled Lapse Rate = $e^{\text{Linear Predictor}}$	38.1%
Actual Lapse Rate Experience	39.1%
Actual Lapse Rate / Modeled Lapse Rate	102.6%

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 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.:   844.64
50-100K :28                       35-39:20  Median :   417.00 Median :  7159.10
GT1M    :28                       40-44:20  Mean   :  1735.03 Mean   : 24594.77
LT50K   :28                       45-49:20  3rd Qu.:  1775.75 3rd Qu.: 24881.78
                                  50-54:20  Max.   :14712.00  Max.   :186853.50
                                  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        65939.43
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.

3. Model Creation

After the basic understanding of the data is obtained, one can start building a model. In the dataset, our target variable is the number of lapses per number of policies exposed per unit of time (in this case, one year).

In this sample model, the Poisson distribution is used and logarithm is the default link function.

The number of lapses is called ‘LapsedN’ in our model and ‘Exposure’ reflects the number of policies exposed for the corresponding duration. To reflect this in the model and since the link function is the logarithm, the offset is the logarithm of ‘Exposure’.

log(LapsedN / Exposure) = log(LapsedN) – log(Exposure)

As we can see from the preceding equation, subtracting “log(Exposure)” on the right side of the equation as an offset is equivalent to dividing by ‘Exposure’ on the left side of the equation, which changes the lapse count to the lapse rate which is what is being modeled here.

```
> Model1 <- glm(LapsedN ~ offset(log(Exposure)) + FaceAmount + PremiumMode + RiskClass +
IssueAge, family=poisson(), data=lapseData)
```

In the above command, “glm” is the specified model family, and ‘family=poisson()’ is the specified distribution. Since the default link function of logarithm is what’s needed, it is not necessary to specify in the bracket. The target variable is ‘LapsedN’, and there are 4 explanatory variables: ‘FaceAmount’, ‘PremiumMode’, ‘RiskClass’, and ‘IssueAge’.

After the model is fit with the data, the model results can be checked with the following command:

```
> summary(Model1)
Call:
glm(formula = LapsedN ~ offset(log(Exposure)) + FaceAmount + PremiumMode
              + RiskClass + IssueAge, family = poisson(), data = lapseData)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-14.4278   -1.7662   -0.1371    1.6875   14.4382

Coefficients:
                           Estimate     Std. Error    z value       Pr(>|z|)
(Intercept)               -2.872380    0.009227     -311.31       < 2e-16 ***
FaceAmount250K-1M          0.063109    0.004443       14.21       < 2e-16 ***
FaceAmount50-100K         -0.171759    0.010461      -16.42       < 2e-16 ***
FaceAmountGT1M             0.078315    0.007342       10.67       < 2e-16 ***
FaceAmountLT50K           -0.333405    0.054839       -6.08       1.2e-09 ***
PremiumModemonthly        -0.413123    0.004736      -87.23       < 2e-16 ***
RiskClassSM                0.061092    0.006221        9.82       < 2e-16 ***
IssueAge30-34              0.105852    0.010531       10.05       < 2e-16 ***
IssueAge35-39              0.207189    0.010053       20.61       < 2e-16 ***
IssueAge40-44              0.301690    0.009946       30.33       < 2e-16 ***
IssueAge45-49              0.398574    0.009955       40.04       < 2e-16 ***
IssueAge50-54              0.474820    0.010132       46.86       < 2e-16 ***
IssueAge55-59              0.537070    0.010541       50.95       < 2e-16 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 18197.4  on 139  degrees of freedom
Residual deviance:  2395.1  on 127  degrees of freedom
AIC: 3461.5

Number of Fisher Scoring iterations: 4
```

The distribution of deviance residuals is displayed in a summary format. The deviance residuals are similar to the standardized error terms.

Following the list of deviance residuals are the predictor variable list, the coefficients and other statistics which have the same format as a standard Ordinary Least Squares (OLS) model.

The deviances of a null model and the current model are stated at the end of the output. The AIC (Akaike information criterion) is also calculated for generic GLM distributions such as the Poisson, Gamma, and Normal distributions. The last line of the output displays the number of iterations of numeric analysis in the GLM algorithm.

After initial iterations of the model, higher orders of covariates and cross-terms need to be considered to account for the significant interactive effects between the predictor variables.

For this particular sample dataset, the cross term between ‘PremiumMode’ and ‘IssueAge’ can be tested to improve the model’s predictive power.

Here are the R script and results:

```
> Model2 <- glm(LapsedN~offset(log(Exposure))+FaceAmount+PremiumMode+RiskClass + IssueAge + PremiumMode:IssueAge, family=poisson(),data=lapseData)
> summary(Model2)
Call:
glm(formula = LapsedN ~ offset(log(Exposure)) + FaceAmount + PremiumMode
             + RiskClass + IssueAge + PremiumMode:IssueAge, family = poisson(),
             data = lapseData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9645  -1.3702  -0.0883   1.0014   5.2205

Coefficients:
                                 Estimate    Std. Error  z value   Pr(>|z|)
(Intercept)                     -2.772184   0.010354  -267.74     < 2e-16  ***
FaceAmount250K-1M                0.063053   0.004444   14.188     < 2e-16  ***
FaceAmount50-100K               -0.182915   0.010468  -17.474     < 2e-16  ***
FaceAmountGT1M                   0.083356   0.007345   11.348     < 2e-16  ***
FaceAmountLT50K                 -0.341975   0.054840   -6.236     4.49e-10 ***
PremiumModemonthly              -0.775494   0.020480  -37.866     < 2e-16  ***
RiskClassSM                      0.060230   0.006220    9.684     < 2e-16  ***
IssueAge30-34                    0.079421   0.012037    6.598     4.17e-11 ***
```

```
IssueAge35-39                     0.149634   0.011510   13.001     < 2e-16  ***
IssueAge40-44                     0.206212   0.011421   18.056     < 2e-16  ***
IssueAge45-49                     0.272848   0.011431   23.870     < 2e-16  ***
IssueAge50-54                     0.321941   0.011663   27.602     < 2e-16  ***
IssueAge55-59                     0.362104   0.012186   29.716     < 2e-16  ***
PremiumModemonthly:IssueAge30-34  0.067422   0.024827    2.716     0.00661  **
PremiumModemonthly:IssueAge35-39  0.188513   0.023577    7.996     1.29e-15 ***
PremiumModemonthly:IssueAge40-44  0.343423   0.023177   14.817     < 2e-16  ***
PremiumModemonthly:IssueAge45-49  0.468724   0.023182   20.219     < 2e-16  ***
PremiumModemonthly:IssueAge50-54  0.578425   0.023481   24.634     < 2e-16  ***
PremiumModemonthly:IssueAge55-59  0.659804   0.024234   27.227     < 2e-16  ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 18197.4  on 139  degrees of freedom
Residual deviance:   427.6  on 121  degrees of freedom
AIC: 1506

Number of Fisher Scoring iterations: 4
```

As seen in the result, by adding the cross term, the AIC is significantly reduced from 3462 to 1506 and residual deviance decreases from 2,395 to 428. The inclusion of the cross term substantially improves our model's performance.

It is tempting to add as many cross-terms as possible to improve the model performance. However, it is important to balance the model fit with both simplicity and business judgment.

A model should be validated to test its effectiveness. There are many techniques available for this purpose; however, they will not be discussed here due to the scope of this brief introduction.

4. Prediction and Result Visualization

After the model is built, the model is then used to predict lapse rates.

```
> lapseData$pred <-predict(Model1, lapseData, type="response")
```

In this command, the model "Model1" is applied to the dataset "lapseData". The prediction is the response of the model, which is the predicted mean value. Other options are available, such as confidence level and uncertainty.

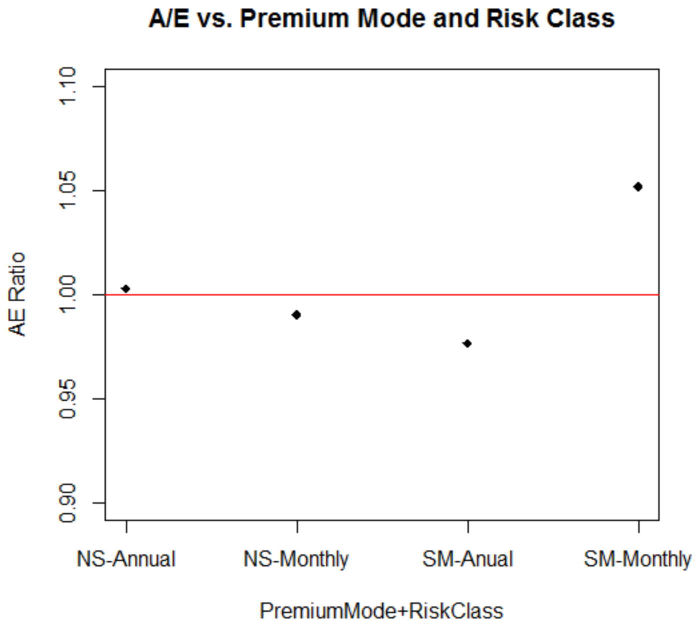
With both predicted values and observed values available, plots can be made to illustrate the model's goodness of fit by comparing the model's predicted lapses to the actual lapses.

R has very strong built-in graphic capabilities. There are numerous packages available for data visualization. It is simple to export the plots to the clipboard or a stand-alone file in popular formats such as .pdf or .bmp. To make an A/E plot, data needs to be calculated and aggregated. In the following example, A/E is calculated by premium mode and risk class.

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

Appendix C: Sample Educational R Code

```
# Data Loading into R

lapseData <- read.csv("C:\\data\\SampleData2014SOAPM.csv", header=TRUE)


#Data Exploration

summary(lapseData)

head(lapseData)

names(lapseData)

dim(lapseData)

tail(lapseData)

aggregate(LapsedN ~RiskClass, data=lapseData, sum)


#Model Building

Model1 <- glm(LapsedN~offset(log(Exposure))+FaceAmount+PremiumMode+RiskClass+IssueAge, family=poisson(),data=lapseData)

summary(Model1)

Model2 <- glm(LapsedN~offset(log(Exposure))+FaceAmount+PremiumMode+RiskClass+IssueAge+PremiumMode:IssueAge, family=poisson(),data=lapseData)

summary(Model2)

anova(Model1,Model2)


#Prediction

lapseData$pred <-predict(Model1,lapseData, type="response")

byPred <- aggregate(pred ~ PremiumMode+RiskClass, data = lapseData, FUN = sum)

byObsv <- aggregate(LapsedN ~ PremiumMode+RiskClass, data = lapseData, FUN = sum)

AERatio <- byObsv[,3]/byPred[,3]

AERatio
```

```
#Data Visualization
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")

#Data Export
write.csv(lapseData,"modelDataFile.csv")
```