SOCIETY OF ACTUARIES

Article from:

Forecasting & Futurism

July 2013 – Issue 7

Predictive Modeling

By Richard Xu

As predictive modeling (PM) draws more and more attention from the actuary community in life and health sections, actuaries may already read many high-level introduction articles and presentations about PM. These materials usually discuss PM background, potential benefits, general requirements, etc., so readers can understand what PM is and what applications are in insurance industry. Seldom did we see an article and discussion about technical aspects of PM, which may be pertinent to actuaries' educational background and their work. However, too many technical details and complicated statistical equations may also intimidate actuaries and turn them away. This article will focus on the mathematical side of PM so that actuaries can have a better understanding, but with a more balanced approach such that readers will get deeper understanding in mathematics terms, but not too many arduous mathematic equations.

The advancement of statistics in the past decades has provided us with abundant choices of techniques that could find their applications in actuarial science. Yet, due to the uniqueness of the insurance business and data structure, we will only find that certain tools are applicable in business, while others are hard to apply in insurance. In this article, I will focus on these techniques that have been proven to be useful in practice.

MODEL BASICS

As was pointed out in the introduction section, there are countless statistical techniques that can be utilized as PM tools in insurance applications. Generally speaking, any statistical model that relies on variables to explain variance of a target variable can potentially be used for the purpose of predicting future outcome. In the language of mathematic terms, we like to build a model as

$$y_i = f(x_{ij}, \beta_j) + \varepsilon_i \qquad \text{(i)}$$

where y_i is called the response variable, dependent variable, or target variable. This is the variable that has been observed in experience and is to be predicted by the model. x_{ij} are called the explanatory variables covariates, input variables, or independent variables. These are variables that have been observed in historical data, and will be observable in the future for the purpose of the forecast. β_j are coefficients to be estimated in the model-building process. ε_i is the error term, which is very important for modeling, but usually not so for prediction, because in most cases we are interested in expected mean values.

TYPES OF MODELS

Linear regression and Generalized linear model

The most common and simplest model is a linear regression model. This is the bread-and-butter model that is taught in almost all colleges, and anyone with an undergraduate degree has probably had at least some exposure to it. The model essentially says the target variable is a linear combination of independent variable(s)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots$$

$$+ \beta_n x_{in} + \varepsilon = \Sigma_j \beta_j x_{ij} + \varepsilon_i \qquad \text{(ii)}$$

To make a valid linear regression in this basic form, several assumptions are needed. A linear relationship between response and explanatory variables is obviously one, but usually this is not a problem. Either the relationship is inherently linear, or it can be well-approximated by a linear equation over short ranges. In addition, the error term ε_i must follow a normal distribution with mean value at zero and a constant variance, i.e. $\varepsilon_i \sim N(0,\sigma^2)$. Other requirements include: y_i is representative of population, observations are independent from each other, and x_{ij} is error-free, etc.

A common method of estimating β_j is the least squares method, in which β_j is chosen such that $RSS = \sum_i(\hat{y}_i - y_i)^2 = \sum_i(\Sigma_j \beta_j x_{ij} - y_i)^2$ is at its minimum, where RSS stands for Residual Sum Square, and $\hat{y}_i$ is the fitted value. There are close form solutions for β_j in matrix form. The other estimation method to find β_j is maximum likelihood in which the product of probability at all data points is at its maximum. Under the normal distribution, it can be proven in mathematics that both estimations will give the same result.

Unless given a very small data set, it is not feasible to build a real model just with pen and paper. You have to rely on computing software to find β_j. The choice of statistical software is quite abundant; options include R, SAS, SPSS, MatLab, MiniTab, etc. In fact, for a very small simple application, one can even use Microsoft Excel's built-in function by selection "Data" -> "Data Analysis," although it has the limit of only 16 explanatory variables. For a large or complicated model, computing software is the only viable choice. Among the actuarial community, the two most commonly used are R and SAS. The R is free software under GNU license, while the latter one is a commercial product. R is unique, not only because it is free, but also because there is a large online community and a core statistics team to support it. You have a wide choice of educational and academic materials about R, and there will never be a shortage of statistic tools in R to build any particular model. As of now (April 2013), there are close to 4,500 packages available on top of the already abundant basic tools that come with the R system, and the number is still growing.

A linear regression is very basic, yet very powerful and efficient. You can easily find a wide range of applications in almost all industry fields. However, you can hardly find any real application in the insurance industry. The main reason is not because of the ignorance of actuaries, but the unique business model and data structure of the insurance industry in which the assumptions of linear regression model are no longer valid. For example, we know the number of claims in a certain group over a period of time is a Poisson distribution where the variance is not a constant, but equal to the mean value. In this case, a linear model can not be used to describe the process why a certain number of claims are observed. Other examples may include claim amount, which follow a Gamma distribution, or mortality rate on binomial distribution.

... THE LINEAR REGRESSION MODEL IS A NATURAL PART OF GLM.

Luckily, the advance of statistics in the past few decades have prepared us with another model called generalized linear model (GLM). As the name indicates, this model is a natural extension of linear model. We can write the model as

$$g\big(E(y_i)\big) = g(\mu) = \eta = \Sigma_j \beta_j x_{ij} \quad or$$

$$E(y_i) = \mu = g^{-1}(\eta) = g^{-1}\big(\Sigma_j \beta_j x_{ij}\big) \quad \text{(iii)}$$

where g(…) is called the link function which links the expected mean value of target variable and the linear combination of independent variable(s).

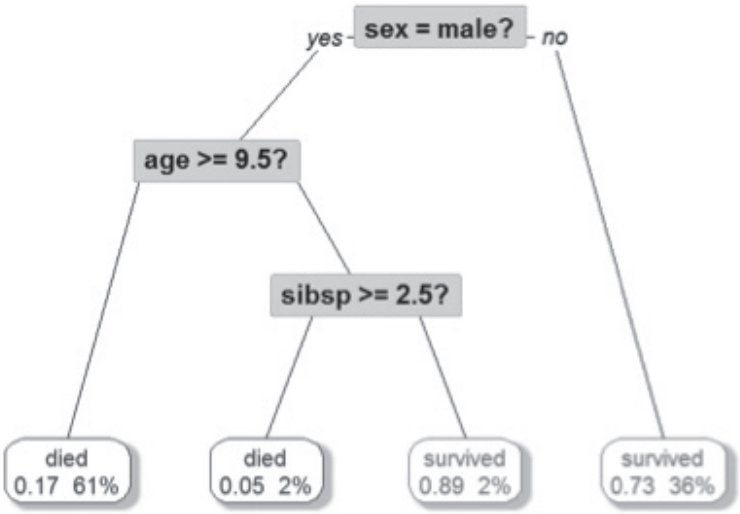
Compared to the linear model, the assumption of normal distribution is no longer needed. Instead, y_i is required to belong to the exponential family of distributions, which is broader and includes most distributions we find in insurance application, such as Poisson, binomial, Gamma, etc. The expansion of distributions also accommodates the variance structure that comes naturally with the distribution. For example, in the Gamma distribution, the variance is proportional to the square of mean. The introduction of the link function makes it possible to drop the strict linear relation between y_i and x_{ij}, resulting with a very flexible model. It is worthy to point out that logarithm could be used as a link function for various distributions. The unique feature of the logarithmic function is that the inverse function is an exponential function such that the additive linear combination in its original form now becomes multiplicative factors. This makes GLM a very powerful tool in the insurance industry as many applications traditionally have multiplicative factors to account for various parameters, such as risk class, gender, industry, locations, etc. Of course, normal distribution is also a member of exponential family, and the linear regression model is a natural part of GLM.

CONTINUED ON **PAGE 14**

Table 1 GLM: link function, variance and application

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

As GLM covers most distributions that are found in insurance and includes various link functions, it is powerful and versatile, and currently is the main focus of PM in insurance. Its applications cover almost all aspects of the insurance business, such as underwriting, actuarial applications (pricing, reserves, experience study, etc.), claims administration, policy management, sales and marketing, etc. Please refer to Table 1 GLM: link function, variance and application.



Decision tree/CART

Besides GLM, another type of model that you may often hear of is an algorithm that is based on a decision tree. In its simplest form, data are split into smaller sections, called leaves, such that data in each leaf will be homogeneous to a certain degree and the variance in data can be explained by a chain of splits on a series of variables. Certain criteria are used to determine which variable to split and at which value so that the split will be optimal.

The most popular decision-tree-based model is the Classification And Regression Tree, also referred to as CART. As the name indicates, you can use this model for both regression and classification. For regression, the target variable is a continuous amount and the model is used to calculate the expected mean value. In this case, the sum of the squares error is used as a criterion to select the split point. In classification, the goal of the model is to separate data into two or more groups. There are several options to accomplish this, such as Gini measure, entropy, etc.

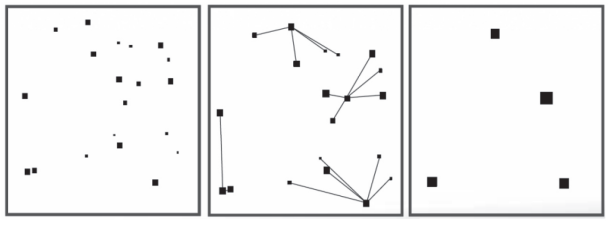
The main advantage of CART model is its intuitiveness and simplicity. When you lay out the tree diagram and present it to your audience, it is very easy to understand and discuss. For example, the Figure 1 A CART model shows a CART model to exaplain the difference of survival rates for Titanic passengers. The decimals at the bottom of each leaf are the probabilities of survival, and the percentages are fractions of the observations. Considering how split variables are chosen and at what value to split, the model itself is quite sophisticated, yet the result is intuitively simple for your audience to grasp the essence of the model without complicated math involved. Other advantages include the non-parametric nature in which you do not have to specify a distribution as assumption, and the automatic handling of possible missing variables. As no model is perfect, the main issue with using the CART model is its low efficiency in dealing with linear relation and its sensitivity to random noise.

Actually, we have already seen this type of model in the insurance business. Think of the process in underwriting, where the information about an applicant will go through an array of decision-making points and finally reach to its final underwriting results. This is exactly the same idea of CART model, although the underwriting processes are built based on experience study and business expertise, not on statistical algorithms. We believe the current underwriting model can be further improved with the help of a decision tree algorithm.

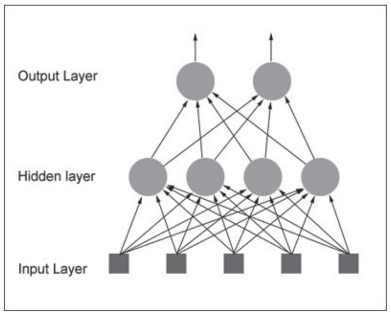
Besides the CART model, there are some other algorithms that are based on the decision tree, but instead of only one decision tree, a group of decision trees are built to extract more information from data. These algorithms are usually much more advanced and sophisticated, but also harder to interpret and gain business insights from. Examples include random forest, and Ada-boosting.

Other models

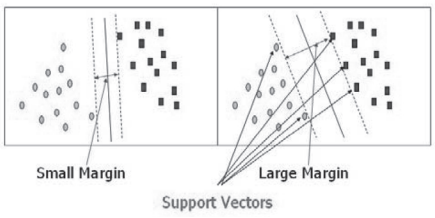
The advance of statistics has brought us more sophisticated models than are discussed above that will potentially find their ways to the insurance applications. Many of them have been utilized in other industries under such names as "business analytics," "big data," or "data mining." Some of them may well be suitable for application in insurance, and a few examples are presented here for illustration.



Clustering. This algorithm is to organize data points into groups whose characteristics in each group share similar distributions. It is an ideal candidate model for applications in classification, especially when the target variable is unknown or not certain. There are many different algorithms to form clusters, but the most popular and simplest is based on Euclidean distance in multidimensional space. You may apply the clustering for market segmentation to find a group of customers that will buy similar merchandises, for identification of effective advertisement for different consumer groups, for recommender systems, etc. In actuarial science, clustering is a very useful tool for in-force cells compression or scenario reduction, especially when a detailed seriatim study is needed or a large number of scenarios have to be simulated.



Neural Network. Also called artificial neural network, the neural network model has its deep root in biological neural networks. The algorithm mimics the interconnected biological neural cells and uses weights for each connection to model patterns in data or relation between inputs and outputs. This model is very powerful in mathematics such that it can replicate any distribution in theory. Its applications date back to the 1990s and today you can find its usage in almost every industry. The neural network is essentially a black-box approach, and it is very hard to interpret the model once it is built. Although its effectiveness and predictive power have been proven in practice, the model cannot help to better understand the business and provide insightful clues to improve it, which limits its application in real business.



SVM. This refers to support vector machine. The basic idea is to split data into two groups in such a way that the separation margin between them is at a maximum. The real algorithm is much more complicated than the

CONTINUED ON **PAGE 16**

simple idea, with multidimensional nonlinear feature space mapping and inclusion of regression as well as classification. This model is generally more accurate than most other models, and is very robust to noise and less likely to have over-fitting problems. Although it is not totally a black-box algorithm, it is still hard to interpret the model and may take a long computing time for a complicated model. Nevertheless, it has great potential in applications in insurance.

The choice of a model that is the best fit to a specific business purpose does not have to be limited to the models that have been briefly discussed here. There are certain rules to follow when selecting a model, but there is also a combination between science and art when you have the freedom to choose between varieties of options. The most advanced and sophisticated model is not necessarily the best choice for a particular business situation. More often than not, some simple models such as GLM may well meet the accuracy requirement and produce desirable results. As long as a model can meet the demand of real business, it will be much more effective to choose a simple model than a complicated one.

CONCLUSION

It should be clear by now that predictive modeling provides a wide range of potential applications for insurance companies. Whether it is a logistic regression model for an underwriting process, a Cox proportional hazard model for a mortality study, or a CART model for pricing, the same core objectives are sought—maximizing the value of data to improve business processes and customer experiences.

To build a successful predictive model that has business value in practice, statistical skill is certainly a very important part of the equation. Actuaries need more education and training in statistical modeling skills. Far too often the statistical nature of the models creates uneasiness for a vast majority of actuaries. On the other hand, mastering of statistical techniques alone does not guarantee a fruitful PM project. Statistical experts often lack the topic specific experience of the businesses for which models will be applied. The merge of these two sets of skills will need a high degree of collaboration between the statistical modeling teams and the business unit experts in order to maximize the potential of PM in business. Actuaries could play an irreplaceable role in the process. ▼

Richard Xu

Richard Xu, FSA, PhD, is senior data scientist and actuary at RGA Reinsurance Company in Chesterfield, MO. He can be reached at *rxu@rgare.com*