SOCIETY OF ACTUARIES
PREDICTIVE ANALYTICS AND FUTURISM SECTION

Predictive Analytics and Futurism

ISSUE 15 • JUNE 2017

Understanding Autoregressive Model for Time Series as a Deterministic Dynamic System

By Dihui Lai and Bingfeng Lu

Predictive Analytics and Futurism

Issue Number 15 • June 2017

Published twice a year by the Predictive Analytics and Futurism Section of the Society of Actuaries.

475 N. Martingale Road, Suite 600
Schaumburg, Ill 60173-2226
Phone: 847.706.3500 Fax: 847.706.3599
www.soa.org

This newsletter is free to section members. Current issues are available on the SOA website (*www.soa.org*).

To join the section, SOA members and non-members can locate a membership form on the Predictive Analytics and Futurism Section webpage at *http://www.soa.org/predictive-analytics-and-futurism/*.

Publication Schedule
Publication Month: December
Articles Due: 9/19/17

From the Editors: A Rainbow of Opportunities

By Dave Snell and Kevin Jones

Welcome! Once again, we have another collection of exciting articles relating to predictive analytics and futurism. Usually, we try to summarize a theme here; but the theme in this issue seems to be "something for everyone," and we mean that in a positive way. The PAF section has become the fastest-growing one in the SOA, and we are proud to say that we have doubled in size over the last couple of years. We would like to think that the newsletter is one of the factors that attracts and retains members; and your excellent article contributions are showing us the vast variety of interests in both the highly quantitative, sophisticated algorithms of predictive analytics and the less tangible but forward thinking aspects of futurism. Before we describe the articles in this issue, though, we want to point out two new structural features:

1. We moved the issue date from July to June (this makes the two semiannual issues six months apart (June and December), and it ought to help make our pipeline of contributions more consistent between the two issues each year.

2. We will be adding a new feature to the newsletter section of the PAF website *https://www.soa.org/sections/pred-analytics-futurism/pred-analytics-futurism-newsletter/* so that you can download an Excel file of all the previous article titles and descriptions along with hyperlinks to the issues that contain them. This will allow you to sort, filter and do all the creative things actuaries use Excel for to manage your library of more than 150 (and growing) articles of interest.

Now, let's talk about what we have for you in this issue:

"Chairperson's Corner: Jump on the PA Bandwagon" by Ricky Trachtman: Ricky claims in his article that English is not his first language, but he is eloquent in describing how to answer his neighbor's question "Don't actuaries predict stuff? You should jump on the bandwagon." Ricky describes how we are both jumping on the PA bandwagon and, in some cases, leading the band.

"Podcasts—A Drill You Can Enjoy!" by Geof Hileman: One area of PAF leadership in predictive analytics has been in the use of podcasts. Geoff describes how you can use these affordable (free), accessible (on your smartphone or similar mobile device) and highly informative audio resources.

"Understanding Autoregressive Model for Time Series as a Deterministic Dynamic System" by Dihui Lai and Bingfeng Lu: In this article, Dihui and Bingfeng describe some of the "art" as well as science in modeling times series data. They introduce a seasonal difference variable—a clever approach that avoids modeling the periodic behavior (more difficult) of these autoregressive models and provides the reader with some guidance for developing an instinct while working with them.

"Predictive Model Building 101" by Dorothy L. Andrews: Dorothy had an ambitious goal for this article—"a guide to help you navigate through 10 modeling phases for building a predictive model and provide you with some insights as to how to overcome obstacles you will likely encounter along the journey." She did it! This is a thorough description of how to get started, how to do it, and then how to validate, test, integrate and monitor it. It's worth saving as a checklist for many predictive analytics projects.

"Predictive Modeling Techniques—A Case Study in Resolving Correlated Explanatory Variables" by Vincent J. Granieri: Following up on his article last year that introduced us to using the Cox Proportional Hazards Model in an underwriting environment, Vince shows how regressing data to find the impact on a dependent variable of many explanatory variables is a worthwhile exercise when building an underwriting debit/credit model.

"Ground Assessment of Soft Skills in Actuaries" by Syed Danish Ali: Danish has written several articles for us, and up until now they were all very technical. He certainly has the technical skills, but this issue shows a softer side to his talents. In this article he quotes Nietzsche: "You must have chaos within you to give birth to a dancing star." We have had several articles on behavioral economics and the importance of the softer, nonquantitative, aspects of predictive analytics, but we think you'll enjoy his unique perspective on it.

"Using Python to Solve, Simplify, Differentiate and Integrate Mathematical Expressions" by Jeff Heaton: We knew Python was cool. But Jeff introduces us to an amazing free and open-source package that allows you to essentially get many of the benefits of Mathematica (expensive) and even have program control over it. Imagine doing symbolic equation simplifications, differentiation and integration within your code with almost no effort. We recommend it!

"On Building Robust Predictive Models" by Mahmoud Shehadeh: Mahmoud takes a deeper look at statistics for us and clarifies how to use a single hold-out validation on a large, publicly available training set (more than 100,000 records). In his article he shows the dangers of using generalized linear model (GLM) results without further investigation of a larger number of samples.

"Speculative Fiction Contest and the Predictive Analytics and Futurism Section Award" by Ben Wolzenski: In case you ever wondered where futurism fits into our section name, Ben describes the recent actuarial speculative fiction contest and lets you know how to get to more than two dozen original science fiction stories written by actuaries and involving some aspects of predictive analytics. They are entertaining and often quite thought provoking.

"Data Visualization for Model Controls" by Bob Crompton: Predictive analytics best practices are not confined to one SOA section; and Bob shares some from his recent article from the March 2017 issue of the *Financial Reporter*. As Bob says in his article, "Can we do better than subject model reviewers to such a painful exercise?" as reading through huge tables of output. He describes several options and points out the strengths and weaknesses of many creative visualization approaches.

"Using Predictive Modeling to Risk-Adjust Primary Care Panel Sizes" by Anders Larson: Anders describes the use of nontraditional techniques for traditional risk situations, specific to health insurance risk scores. He shows the advantage of an ensemble of smaller models utilizing gradient boosting machines as an alternative to the more traditional generalized linear model. Along the way he describes using cross-validation to avoid overfitting, and Anders concludes, "There is not a one-size-fits-all solution to risk adjustment."

"Bayesian Inference in Machine Learning" by Denis Perevalov: Just because maximum likelihood estimations (MLEs) are fast and scalable does not mean they are the best choice in all machine learning situations. Denis takes us back to basics with Bayesian inference, which he shows may be a better choice for smaller amounts of data or data that are narrow in the longitudinal direction. Plus, he shows that Bayesian inference can make more precise predictions, and its confidence intervals of model parameters are more interpretable.

"Maximal Information Coefficient: An Introduction to Information Theory" by Bryon Robidoux: Before reading Bryon's article, we might have assumed that a "bit" is a binary digit. This is unfortunate because a binary digit and a bit are different, and Bryon shows how in this summary of information theory. Along the way, he also introduces us to "nats" and "bans" and conditional entropy. Pattern matching is a lot more mathematical than we supposed, and we can draw upon information theory for a head start.

"Variable Selection in Predictive Modeling: Does It Really Matter?" by Kailan Shang: Are more variables always better when you are trying to build a predictive model? Kailan tells us no! In addition to the added complexity, you run into the presence of collinearity caused by too many variables, and this detracts from the robustness of your models. Kailan shows ways to reduce the number of variables and stresses the need to use human expert judgment at various stages of the process.

"The First SOA Annual Predictive Analytics Symposium—A Recommended Investment! (Whether or Not Your Employer Pays for It)" by Dave Snell: The section growth and the burgeoning interest in predictive analytics have reached a new milestone. We are planning to have our own annual special interest meeting, like the Health section, the Life and Annuity folks, and the Valuation actuaries. Dave describes this new SOA event and explains why you should sign up for it just about the time you get this newsletter. Don't wait! Sign up now!

Enjoy the issue! ■

Dave Snell, ASA, MAAA is technology evangelist at SnellActuarialConsulting in Chesterfield, Mo. He can be reached at *dsnell@ActuariesAndTechnology.com.*

Kevin Jones, FSA, CERA, is associate actuary at Milliman in Buffalo Grove, Ill. He can be reached at *Kevin.Jones@Milliman.com.*

Chairperson's Corner: Jump on the PA Bandwagon

By Ricky Trachtman

A couple of weeks ago I was having a conversation with a friend. We do not usually talk about what we do at work, but she knows that I am an actuary. She told me that she read an article about big data. She mentioned that the article said that one of the most useful things about big data was the ability to predict outcomes. Then she said something like "Don't actuaries predict stuff?" and then followed by saying, "You should jump on the bandwagon." I don't know why, but I felt a bit hurt by her comment.

As some of you may know, English is not my first language, but I knew what the gist of the phrase was. For those not familiar with the phrase, "jump on the bandwagon" means to begin supporting or following a hobby, idea, person, method etc. after it has become popular or successful. I did not know why the phrase used a bandwagon versus any other wagon. So on one of those internet breaks one needs to have, I researched the phrase. I found that the word "bandwagon" itself is the name for a wagon that carried a band. These wagons were bright and ornamental, and they were almost impossible to miss. They were in front of parades, circuses and political rallies to make a big entrance and to say, "Here we are!"

At some point next year, actuaries will have as part of their associateship required education instruction on predictive analytics. The next generation of actuaries will be required to not only understand statistical techniques, but also to use them with tools that they can apply directly at their jobs. There is an interesting article, by Stuart Klugman, in *The Actuary* magazine that came out in February/March of this year titled, "Put to the Test," in which he explains some of the complexities of testing the material. For the rest of us, to use predictive analytics we need to self-learn.

There are some challenges with self-learning, but most of us had to do it to pass the required examinations to become actuaries, which in one form or another make us familiar with the process. The Predictive Analytics and Futurism (PAF) Section has striven to provide you with many resources to learn predictive analytics at many different learning levels. Our very popular podcasts are an excellent way to keep learning and improving your knowledge. We sponsor multiple sessions through the symposiums and meetings from the SOA throughout the year on the subject. We will have our second practical predictive analytics seminar after the Life and Annuity Symposium. In this hands-on seminar, participants learn how to build a basic predictive model using R. In September of this year, the SOA will debut its first annual Predictive Analytics Symposium, which Dave Snell describes in another article in this issue. This newsletter is a wealth of knowledge for all levels of learning. I encourage all of you to not only read this issue, but also to explore all of our prior issues, because those articles continue to be relevant and highly informational.

If you are just starting on your predictive analytics self-learning journey, or know of someone who is starting, a vast number of courses and books are available, many of them free, that will aid you on your way to knowledge. I would encourage you or them to read the article from Mary Pat Campbell, "Getting Started in Predictive Analytics: Books and Courses," from our December 2015 newsletter issue. It truly is a great place to start. Even if you think you are not going to build predictive models, I would encourage you to learn some of the tools and computer languages used to build predictive models. These tools are very useful to understand and process information/data that you may already have. These tools may provide ways to analyze and view results in different ways, allowing us to gain a better understanding of the information. There is a great group of research articles on the SOA website about data visualization in actuarial practice that may help you spark ideas on new ways to view information.

Predictive analytics does seem very much like a bandwagon for an actuary nowadays. It appears in publications, meeting sessions, webcasts, podcasts—you name it. It is right there, bright, loud and ornate. As my friend said, we actuaries do predict "stuff" and we do much more. We have been predicting using data for a long time, but now we can do it utilizing exciting new tools, new sources of information, new techniques, and we can do this at a more granular level than ever before. These are exciting times for our profession, and come to think of it, I should have not felt hurt, I should feel proud to be at the forefront of it all, right on top of the PA bandwagon. ■

Ricky Trachtman FSA, MAAA, is a principal and consulting actuary at Milliman. He can be reached at *ricardo.trachtman@milliman.com*.

Podcasts—A Drill You Can Enjoy!

By Geof Hileman

Did you know that you can know earn continuing education credits by going to the dentist? While out for a run? While cutting your hair? Thanks to the SOA's relatively new and rapidly expanding library of podcasts, you can earn CE just about anywhere you choose.

For the uninitiated, a podcast is simply a free audio recording that can be downloaded and played back on any computer, tablet or smartphone. Podcasts have been growing in popularity at such a rapid rate that it is difficult to find current statistics. One estimate shows that more than 57 million Americans listen to podcasts at least monthly. The appeal of the podcast medium mirrors the something-for-everyone nature of other web-based media. Whatever your interests, there is certain to be a podcast just for you. I've listened to podcasts ranging from a 12-hour series on the thirteenth-century expansion of the Mongol Empire to a 20-minute discussion of why milk is placed in the back of the grocery store (the answer is as obvious as you'd think).

While perhaps holding less universal appeal than some other more frequently downloaded podcasts, readers of this newsletter may find common ground in the podcasts that are now available through the SOA and, in particular, those produced by the Predictive Analytics and Futurism Section. All of the SOA's podcasts can be found by searching any podcast provider for "Society of Actuaries Podcast Feed." Within the search results, you will find a list of podcasts created by individual sections and by the SOA generally. As with podcasts in general, the SOA podcasts cover a wide range of actuarial subject areas.

The PAF podcasts are hosted by Anders Larson and Shea Parkes, both PAF Section Council members. Each episode is 15 to 25 minutes in length, and many are part of a series introducing listeners to various predictive modeling techniques and topics. Previous episodes have discussed the bias-variance tradeoff that lies at the heart of many predictive modeling exercises, cross-validation and bootstrapping, penalized regression, random forests, decision trees and ensemble modeling. In addition to the series covering specific modeling topics, the podcast has also discussed other special topics. One recent episode featured an interview with leaders of a firm that has been expanding its predictive modeling capabilities and discussed the approach they had used to build out and deploy those new capabilities.

Using podcasts as one component of meeting your continuing education requirements offers three distinct benefits. First, it's difficult to beat "free," especially in the realm of actuarial continuing education. Second, the convenience is unparalleled. Episodes can be downloaded in seconds and listened to wherever you go. Third, and most importantly, the podcast vehicle provides access to leading experts providing highly current information. Even conference presentations and newsletter articles are planned months in advance, but podcasts can be put together very quickly and can provide actuaries with highly relevant and recent information.

At the risk of damning with faint praise, I can personally report that the PAF podcasts are far better than the sound of the dentist's drill. Better than that, the SOA's library of podcasts are a great new way to stay current and to help meet your continuing education requirements. ■

Geof Hileman, FSA, MAAA, is VP at Kennell and Associates Inc., in Raleigh, N.C. He can be reached at *ghileman@kennellinc.com*.

Understanding Autoregressive Model for Time Series as a Deterministic Dynamic System

By Dihui Lai and Bingfeng Lu

The autoregressive (AR) model is commonly used to model time-varying processes and solve problems in the fields of natural science, economics and finance, and others.[1] The models have always been discussed in the context of random process and are often perceived as statistical tools for time series data. However, randomness is only part of the story. The rich deterministic dynamics that an AR model produces is perhaps also worth some attention.

In this article, we are going to discuss the AR model by making connections to time-dependent ordinary differential equations. The goal is to understand the essential dynamics underlying the AR model and provide guidance on model usage in addition to statistical diagnostic tools.

AUTOREGRESSIVE MODEL

In general, the autoregressive model describes a system whose status (dependent variable) depends linearly on its own status in the past. The system can be mathematically described by a stochastic difference equation such as the following:

$$y_t = \beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i} + \varepsilon_t .$$

Here, the βs describe how much the system's status i steps ago will impact current values. Normally, one would expect βs to decrease as i increases, that is, the events that happen further in the past have less impact on current events. Anything that happens earlier than p time steps ago will have no impact, and the model is noted as AR(p), where ε_t is a "noise" term that describes some random events that affect the status of the system. The "noise" term is often required to be stationary to make lots of statistical estimators valid (least-square estimation, maximum-likelihood estimation etc.).

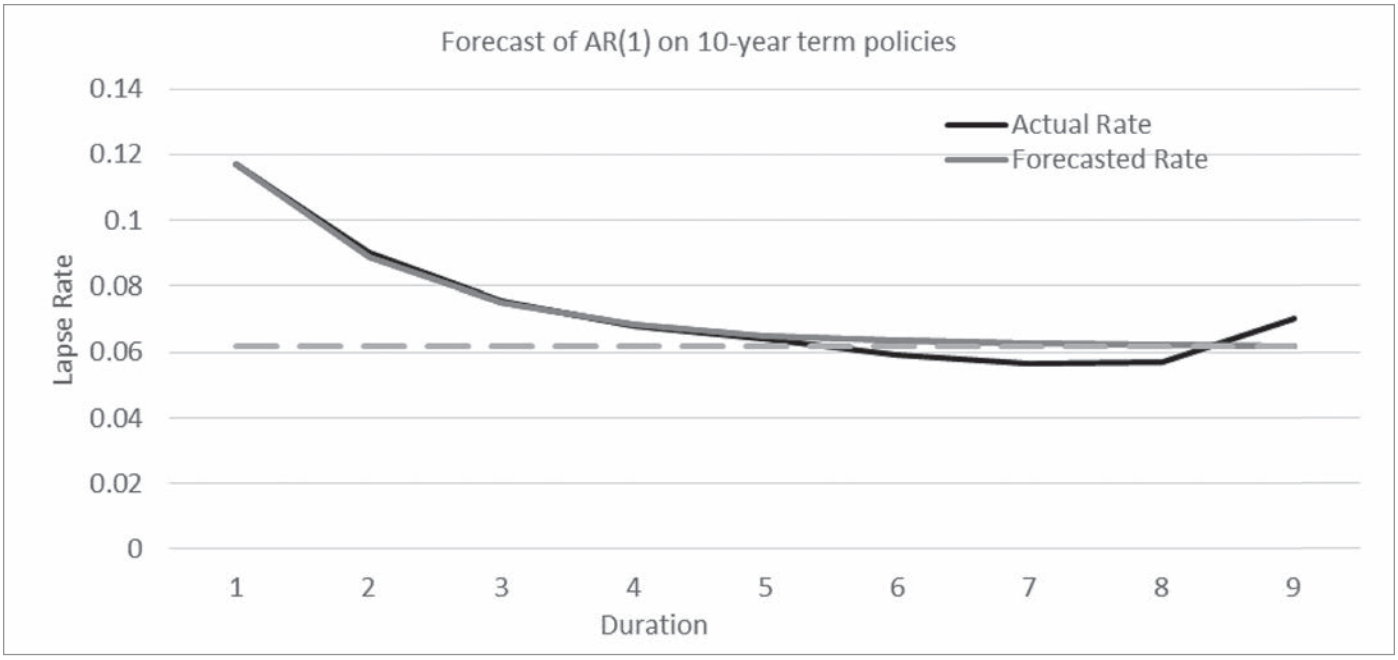
AR(1) MODEL AND FIRST ORDER TIME-DEPENDENT ORDINARY DIFFERENTIAL EQUATION (ODE) SYSTEM

In a very simple scenario where $p = 1$, we have an AR(1) model where the system's current status is dependent only on the system's status one time step ago: $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$. The continuous version of the system can be represented as a first-order time-dependent ODE with a noise term: $\frac{dy}{dt} = \beta_0 + (\beta_1 - 1) y + \varepsilon_t$ (see the appendix). Without considering the noise, the closed formula solution of the ODE is an exponential function: $y = constant * e^{(\beta_1 - 1)t} + \frac{\beta_0}{1-\beta_1}$. It follows immediately that the status of the system will reach an equilibrium point $\frac{\beta_0}{1-\beta_1}$, if $\beta_1 < 1$, as the exponential term vanishes in the long term. Not surprisingly, this is also the expected behavior of an AR(1) model in equilibrium status when $y_t = y_{t-1}$.

Now that we have made the connection between the two systems, it becomes clear that the parameter $(1-\beta_1)$ could be interpreted as a decay constant that describes how fast the system will reach a steady value as time elapses. When $\beta_1 < 1$, the AR(1) model is nothing more than a system that exponentially decays to a steady state from a certain initial value noted as *constant* in the close formula solution. On the other hand, when $\beta_1 > 1$, the dependent variable will exponentially increase to a very large value.

In another words, an AR(1) model can be used to describe the evolvement of systems that have decay-like behavior with a long-term equilibrium point. As an example, we modeled the lapse behavior of a 10-year term life policy over the level period with an AR(1) model. The model uses the lapse rate at each policy year as the target variable. To make a forecast, we provide the model with an initial lapse rate at duration 1, and the lapse rate evolves as an exponential decay toward a stable point (see Figure 1). The model forecast did quite well at early duration but underestimated the rate after duration 5, indicating that extra factors need to be considered beyond the dynamics described by AR(1).

Figure 1
Forecasting the lapse rate of a 10-year term life policy over a level period by the AR(1) model. The black line is the actual lapse rate, and the red line is the forecasted rate. The forecasted lapse rate quickly decays and reaches a stable point (green dashed line)



AR(*p*) MODEL AND *p*TH-ORDER TIME-DEPENDENT ODE SYSTEM

In general, an AR(*p*) model is a *p*th-order linear difference equation with a noise term. It can be proven with some linear algebra techniques that a *p*th-order linear difference equation can be reorganized into a set of *p* first-order ODEs. Thus, it is expected that an AR(*p*) model will inherent some dynamic properties of a *p*th-order ODE set. In the following section, we use an AR(2) model to reproduce the behavior of an oscillatory system.

SEASONALITY OR HARMONIC OSCILLATOR?

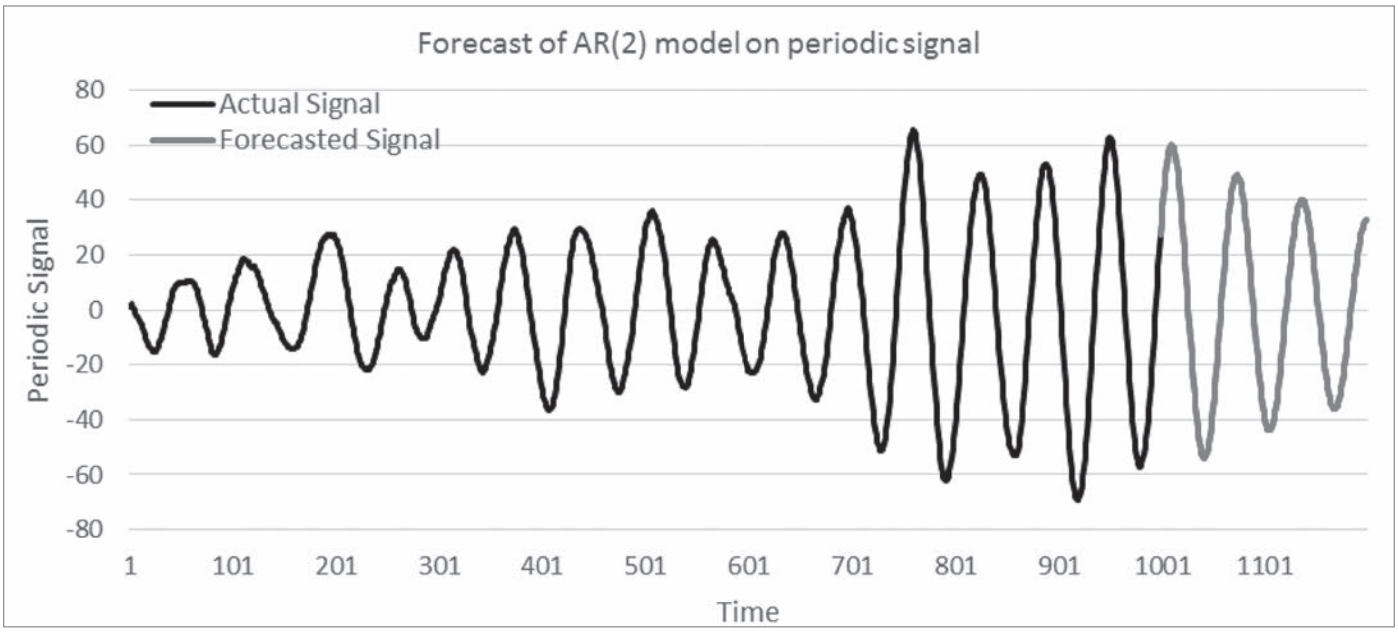
When studying time series, the periodic behavior is commonly modeled by constructing a new seasonal difference variable $\Delta y_t = y_t - y_{t-T_{period}}$. The evolvement of the system over time is then described by the new variable Δy_t. This clever approach avoids modeling the periodic behavior by removing the gross seasonal feature and considering only the change over seasons.[2] However, to make a forecast, this approach needs to have n initial condition parameters where $n = T_{period}$ and some prior knowledge for T_{period} are needed.

Alternatively, we know that a second-order ODE system will lead to oscillatory behavior (a harmonic oscillator can be described by a second-order ODE) given the right parameter sets, and therefore we expect the time series version of the system will produce periodic behaviors. As a demonstration, we build an AR(2) model on a sinusoidal time series signals (see Figure 2). Without explicitly modeling the seasonal activity, the model captures the essence of the oscillatory behavior (period) with only three parameters.

SUMMARY

When building an autoregressive model, it is often more of art than science to decide the value for *p*—that is, how far do we have to trace the system's past to make a reliable forecast? Some tools are available to help the decision-making process, such as an autocorrelation function (ACF) or a partial autocorrelation function (PACF).[3] Although the diagnostic tools provide convenient guidance on choosing the lag parameter, it is not always easy to find a clear-cut value. The judgment becomes even harder for a noisy data set.

Figure 2
Forecast of an AR(2) model on a periodic signal. The black line is the original signal, and the red line is the forecasted behavior of the system. The forecasted part reproduces the periods of the signal quite well.



In this article, we demonstrate the dynamic feature of AR models. By borrowing concepts and closed formula solutions from time-dependent ODEs, we gain some intuition for the parameters in AR models (βs and p) and relate them to the dynamic properties of continuous systems. We use some examples to demonstrate that an AR(1) model can be used to model a dynamic system showing decay-like behaviors. Besides the commonly used seasonality model, an AR(2) model could be used to model periodic oscillatory (seasonal) behaviors. ■

Dihui Lai, Ph.D., is a senior data scientist at RGA Reinsurance Company in Chesterfield, Mo. He can be reached at *dlai@rgare.com*.

Bingfeng Lu is an assistant data scientist at RGA Reinsurance Company in Chesterfield, Mo. He can be reached at *blu@rgare.com*.

ENDNOTES

1. Zhang, Yuanjie Michael, Jeffrey R. Russell and Ruey S. Tsay. 2001. "A nonlinear autoregressive conditional duration model with applications to financial transaction data," *Journal of Econometrics* 104, no. 1, pp. 179–207.
2. Hyndman, R.J. and G. Athanasopoulos. 2014. *Forecasting: Principles and Practice.* Melbourne: OTexts.
3. Box, G. E. P., G. M. Jenkins, G. C. Reinsel and G.M. Ljung. 2015. *Time Series Analysis: Forecasting and Control.* 5th ed. Hoboken, NJ: John Wiley & Sons.

APPENDIX

First-Order Difference Equation and First-Order ODE

A first-order difference equation can be written as $y_t = \beta_0 + \beta_1 y_{t-1}$. Here we have assumed a change over one time unit in the formula. In general, the time step can be of any unit, and by changing the unit of time, we can replace unit time with Δt, and the equation can be rewritten as $\frac{y_t - y_{t-\Delta t}}{\Delta t} = \beta_0 + (\beta_1 - 1) y_{t-\Delta t}$. When $\Delta t \to 0$, the difference equation becomes a first-order time-dependent ODE $\frac{dy}{dt} = \beta_0 + (\beta_1 - 1) y$.

Predictive Model Building 101

By Dorothy L. Andrews

Your boss has just given you a project to build a predictive model to identify highly profitable customers, and you have no idea where to start. The predictive modeling exercise begins with understanding the business problem and ends with validation of the model and dissemination of the results. However, there will be the ongoing task of monitoring the model for continued fit as new data emerges and the business changes. A lack of fit is a clear signal the model needs either a "refresh" or a "rebuild." You will need to overcome many obstacles in getting a model from the drawing board to company production systems. This article is intended as a guide to help you navigate through 10 modeling phases for building a predictive model and to provide you with some insights as to how to overcome obstacles you will likely encounter along the journey.

PHASE 1: DEFINE THE PROBLEM

The financial objectives of the organization should be a guiding light in defining problem statements your model will address. Management are more likely to allocate resources and sponsorship to your modeling project if the solution addresses "pains" that keep them up at night. If you cannot clearly articulate how your model is important to the continued health of the organization, it is unlikely management will leverage scarce resources to fund its execution. Make sure you define the problem in terms your stakeholders will understand. It is important that management who can eliminate obstacles that may hinder the successful implementation of your model project be included among your stakeholders.

It is important to demonstrate that the problem you wish to solve is observable, measurable and subject to classification on some metric. For example, observable characteristics of a highly profitable customer are the types and number of insurance products they own. However, merely owning a product is not sufficient. We need to measure characteristics such as policy retention, premium payment levels and cancellation/renewal behavior to refine profiles of highly profitable customers. Once profitability criteria are identified, then customers can be rank ordered on a scale from least profitable to most profitable. Management is then better positioned to remediate the least profitable and improve retention efforts to keep the most profitable and find more like them. It is important to keep the financial objectives of the company in mind as you develop your problem statement.

PHASE 2: DEMONSTRATE THE FINANCIAL IMPACT OF THE SOLUTION

Key stakeholders in your organization include members from the C-Suite and senior leaders in the actuarial, underwriting and information technology (IT) groups. Agents and brokers may also be stakeholders since they assist their clients with purchasing products using customer scores resulting from predictive models. For example, if agent portals are equipped to render a customer profitability score based on data entered by the agent, then agents may be motivated to produce the best score possible. Data controls will need to be in place to identify when possibly conflicting combinations of data may adversely impact a customer profitability score.

The proposed model should be of financial significance for each of your stakeholders. It is important to understand how the model will improve the financial position of the organization. The more significant the financial impact, the greater the likelihood your stakeholders will support the implementation of your modeling project.

PHASE 3: UNDERSTAND THE PRODUCTS

Model building begins with a solid understanding of the design and features of the products being modeled, how they are marketed and their distribution channel, and the accuracy of underlying administrative and other company data. Many company administrative systems lack adequate controls around the data entry of application and product attribute data. As a result, it becomes essential for the modeler to develop assumptions regarding missing and incorrectly specified data. This requires expert knowledge of the product's distribution, marketing, features and design. Such expert knowledge is also invaluable in understanding anomalous data elements. For example, if a particular product feature appears more frequently in your data set than it should, then it is important to investigate such an anomaly to determine its validity. The results of the

investigation can often become a teachable moment to improve the administration of application and product attribute data. The fewer assumptions needed to prepare data for modeling, the more reliable the results of the model to measure phenomena of interest to the company.

PHASE 4: IDENTIFY INTERNAL AND EXTERNAL DATA

A number of considerations are necessary when constructing modeling data sets from internal company data. Most financial data is transactional in nature, requiring extensive coding to summarize it, recognizing canceled and backdated transactions, in particular. Failure to recognize the cancellation of premium payments, for example, will lead to overestimating net premiums paid on a policy, impacting any derived metric based on premiums. Many companies still currently rely on legacy systems that require Job Code Language (JCL) and COmmon Business Oriented Language (COBOL) to extract data needed for modeling. Further, the number of programmers familiar with these languages is dwindling, putting a premium on sought-after resources for your project.

Changing IT platforms can be extremely expensive, but most companies recognize the need to make the transition to more relational architectures, and they are making the investment. These architectures need to be more flexible, however, to accommodate the codification of new data elements. For example, it is common that the only data element that captures height and weight is the adjuster note. The adjuster note is an example of an unstructured data element. It is free-flowing text entered at the discretion of the adjuster. These notes represent data-mining gold if you are studying the relationship of height and weight to the duration of workers compensation claims. Although text-mining tools are available to assist with the mining of adjuster notes, companies can gain greater leverage from their data by structuring the collection of data elements once it becomes clear they have predictive value.

Once internal data has been structured, appending external data can significantly increase the predictive power of predictive models. What external data should you include? Good question, because we have lots to choose from. Currently, models are including census data, geospatial variables, economic data and consumer attribute data marketed by companies like Acxiom to assist with customer segmentation. Companies recognize the need to market differently to Millennials than to Gen X'ers and Baby Boomers, and they are incorporating marketing data in their predictive models. Depending on the purpose of your model, it is very important to make sure model results based on internal and external data do not unfairly discriminate against policyholders. Regulators have, as one of their primary missions, to prevent unfair discrimination in the pricing and distribution of insurance products. They are becoming educated on advanced modeling techniques, and they especially scrutinize model variables for their unfair discriminatory power. Do yourself a favor and make sure your in-house counsel reviews your variables, especially if your models need to be disclosed in regulatory filings.

PHASE 5: ITERATIVE DATA SCRUBBING AND ANALYSIS

Modelers are fairly united in their view that most of the heavy lifting in building a predictive model involves scrubbing and analyzing the raw data and augmenting these data with relevant external data. Insurance company data, like that of others, is transactional by design. Every time a change is made to some aspect of a policy, a new data record is created in every company system where the change applies. The first step in constructing the modeling data set is the extracting of raw data from company systems and summarizing these data to an appropriate level and at an appropriate periodicity. For example, data may be summarized at a policy level for every quarter in the model study period. This means your data set contains a snapshot of the policy at every quarter end for the model study period. This is a programming task that is often achieved with the help of the IT department or, what is becoming more likely, by the modeling team to avoid delays often associated with IT project scheduling. When the modeling team takes on this task, it is paramount that control totals are identified to validate modeling data against to assess the accuracy of the programming results. External data is usually appended to the summarized data records.

Missing data and misspecified data are unavoidable in any data set, but if improperly resolved, the data set will likely bias your results in unwanted directions. Resolving missing and misspecified data requires a solid understanding of how the products being modeled are distributed, designed and marketed to develop assumptions and adjustments to transform "messy" data into usable data. Construct frequency distributions of the levels for each attribute variable and histograms for continuous variables as a first step. Discrete variables are often treated as attribute variables if there are a limited number of values in their range. External data can be missing and misspecified if out of date. For example, if policy zip code data is invalid, it may not be possible to append census data, such as average income or home values, two important attributes in life, health and P&C modeling.

Misspecified data elements can be harder to detect. Examining frequency distributions can shed light on values that don't belong in a field. Conducting inspections on dependent fields is also another tool to identify misspecifed fields. Data dependency in this context means the values on one data elements limit the possible values on the dependent data element. The results of such inspections can be used to correct company processes responsible for misspecified data elements. Modelers should feel some responsibility to influence the correction of data anomalies companywide and not just for the modeling exercise. The modeler can use the results of analyzing missing

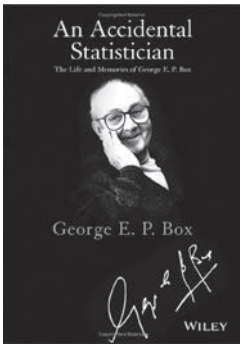
and misspecified data to develop eligibility criteria for including records in modeling data sets. It is, additionally, important to quantify the financial impact of excluded records by some standard of materiality. Modelers may want to exercise more due diligence in correcting records determined to be financially material.

PHASE 6: MODEL VARIABLE DEVELOPMENT

The raw data records have now been cleaned, appended with external data and assessed for inclusion in the final modeling data set. However, the modeling data set may not be complete. Additional considerations include the development or grouping of levels on attribute variables, the derivation of new variables from the raw data, the treatment of variables either stochastically or deterministically and the identification and derivation of a target variable, if applicable. These are nontrivial considerations and a function of the purpose of the model.

It is important to understand the qualitative relationships among the variables in the data set to eliminate variable dependencies. Your predictor variables should be mathematically independent. Examine any correlations that may exist among your variables to avoid including variables that measure the same model effect. A principal components analysis is a useful technique for isolated uncorrelated variables. Examine the clusters of correlated variables to determine which **one** from each cluster to include in the final modeling data set. Correlated variables can lead to unstable parameter estimates and should be avoided in constructing the final modeling data set. Naturally, this extends to derived variables and the variables used to create them. A simple correlation matrix can assist with this identification. Univariate analyses are also useful to identify variables to include, but not the ultimate criteria by which to select model variables. Stepwise procedures additionally can help demonstrate the statistical importance of variables in the presence of other model variables and are yet another tool for finalizing the final set of modeling variables.

PHASE 7: MODEL CONSTRUCTION



This is the phase of the project every modeler loves to reach. This phase involves selecting the "right" statistical model to fit the data. I want to caution modelers in thinking they have the "right" model when they are done with the exercise. In the words of Dr. George E. P. Box, "Essentially, all models are wrong, but some are very useful." Dr. Box founded the Statistics Department at the University of Wisconsin at Madison. He taught himself statistics while serving in the British Army. During that time he became very good friends with Dr. R. A. Fisher, considered to be the founder of modern-day statistics, and he went on to earn a Ph.D. in statistics from the University of London. He is considered to be "one of the greatest statistical minds of the 20th century."[1] Dr. Box co-invented the Box-Cox Power Transformation used in regression analysis with Dr. David Cox, noted for his contributions in the area of proportional hazards regression modeling. Please read Dr. Box's memoir, *An Accidental Statistician: The Life and Memories of George E. P. Box*. You will find it thoroughly captivating and inspirational.

The notion of a "useful" model should remind modelers that a more useful model may exist. Software packages are greatly simplifying the identification of "useful" models using just a few keystrokes. Once the modeling data set has been constructed, software packages are available that will run several kinds of statistical models against the data set and rank order the resulting models under a set of tests of statistical significance. These software packages require little to no program skills to run, but let's face it, running models falls in the 20 percent of the effort category of the "80–20 Rule" as applied to building a predictive model. The real modeling building takes place in transforming the data under a set of modeling assumptions and developing the criteria for selecting potential data variables, which is the 80 percent of the "80–20 Rule." The number of lines of programming code needed to program a generalized linear model (GLM), for example, is a mere fraction of the amount of code needed to build the modeling data set, unless your data is naturally perfect. Naturally perfect data is a modeler's dream, but seldom encountered.

A word of caution is in order in respect to some of these packages. While they may be child's play to use in terms of simplicity, interpreting model results should be left to a subject matter expert with a thorough understanding of statistics, the products being modeled and the business environment in which model results will be applied. Further, don't underestimate the need to clearly articulate model results to your stakeholders. It will be important to demonstrate how the model results solve the proposed problem in terms they understand so they may comment on the model. All your hard work will have been for nothing if you express you results in esoteric statistical jargon your business leaders can't understand, which may compromise the likelihood of its adoption by the company.

PHASE 8: MODEL VALIDATION AND TESTING

Most would agree that recognizing the "wrong" model is easier than qualifying the "right" model, if a "right" model is even possible to build. Model validation can help you assess whether your model is a reasonable representation of the phenomena under study. But remember, the model is only a representation of the "real thing" at a given point in time. It is **not** the "real thing." (Sounds like an ad for Coca-Cola, right?) The phenomena under study is constantly changing, while the models are always in catch-up mode in their predictive power. The greatest flaw of any model is the model risk they pose for organizations using them.

In a 1996 Goldman Sachs "Quantitative Strategies Research Note," Goldman Sachs defined model risk as "the risk of loss by using a model to make financial decisions" and identified several forms of model risk. They identified the following types of model risk: 1) inapplicable model, 2) incorrect model, 3) correct model, incorrect solution, 4) correct model, inappropriate use, 5) badly approximated model, 6) software and hardware bugs and 7) unstable data. The reader is directed to this paper for the details of each type of model risk. However, the meaning of each type of risk should be fairly intuitive. The paper also goes into considerable detail enumerating the signs a model may be incorrect. For example, the modeler may not have considered important factors in the design of the model or the model may be correct only under ideal conditions, which rarely present themselves.

Insurance companies might borrow a page from banking to establish a formal model validation process for vetting company models. In banking, a model validation group is a group of interdisciplinary academics and banking professionals familiar with the company's products and business functions that convenes to vet proposed models before they are presented to senior management. The model validation team, by design, is an interdisciplinary team of professionals who can assess the impact of the model on all aspects of a company's operations, from its distribution channels to its marketing and underwriting departments and processes. The rigorous nature of the validation process is critical to mitigating model risk by identifying weaknesses in models and recommending remedies to increase the likelihood of their company adoption or recommending the nonadoption of models that could adversely harm the company financially. This can be an unpleasant experience for the modeler, but the continued health of the organization is the paramount concern to all involved in the model validation process.

PHASE 9: SYSTEM INTEGRATION

It probably does not come as a surprise that you will need to build a model to test the implementation of your predictive model by the company IT department. The testing of the implementation needs to include enough scenarios to ensure the model behaves as expected once in production. Otherwise, a very soundly constructed model could get a "bad rap" because IT implementation failed to properly operationalize it. In the testing of the IT implementation, don't ignore even the smallest of discrepancies. A seemingly immaterial difference could yield unexpected results once a model goes into production and attempts are made to evaluate a combination of policy data not represented in one of your modeling test scenarios.

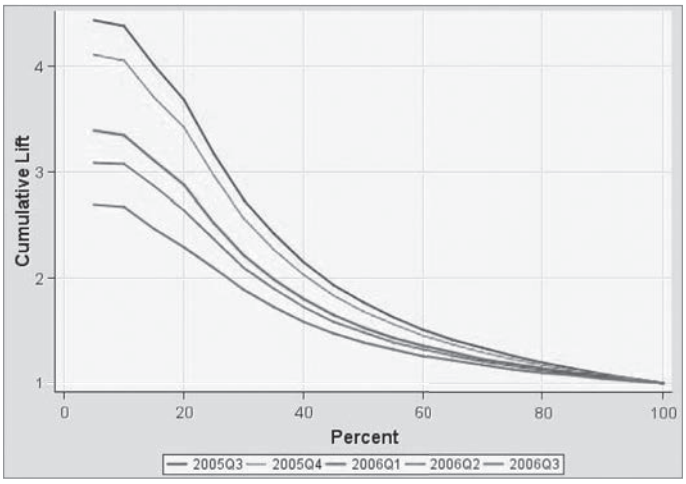
Production models should be tested for their ability to replicate the results of all test scenarios, which should include simple and complex test cases as well as boundary or extreme cases. It can't be stressed enough that the importance of models accurately replicated the simple cases. Models quickly lose credibility with end users and senior management if they fail at replicating simple cases, casting doubt on results for more complex cases. End users then become engaged in scrutinizing model results rather than looking for emerging risks that may challenge the profitably of the organization. When underwriters, for example, spend an inordinate amount of time trying to disprove model results they don't trust, they are engaging in the wrong kind of behavior for the organization. The simple truth is if they don't trust the results, they are not going to use them to make underwriting decisions anyway. The time spent earning "scout badges" every time they disprove a result from the model could have been better spent on behalf of the organization looking for emerging risks. This is a prime reason the model validation exercise is so important. The better the interdisciplinary review of the model and the testing of its IT implementation, the higher the confidence level around the organization for the model and the greater its utilization in decision making.

PHASE 10: DEVELOP MONITORING METRICS

Monitoring metrics are used to assess the continued fit of the model as new data emerges and the business environment changes. If model results are not as expected and/or major distributional shifts from modeled data present in emerging data, then it is time to consider whether the model requires a "refresh" or a "rebuild." Minor distortions may necessitate only a model refresh. A model refresh is performed by running the same model against an updated modeling data set to update model parameters. Major distortions necessitate a complete overhaul of the existing model, which includes developing an entirely new data set based on new model predictor variables. Some of the old variables may still apply, but the degradation of your model is a suggestion they are failing to capture new signal-affecting business metrics.

Chu et al. (2007) discuss many best practices for monitoring predictive models once they have been installed into production. They discuss developing performance thresholds and the automation of the periodic generation of performance metrics to identify when models are underperforming. A key performance degradation tool the authors discuss is the model degradation lift chart exhibited below. A lift chart measures how well predicted values line up with actual values. In this chart, the model is run quarterly to examine how the lift changes over time. One could run the analysis at a frequency greater than quarterly depending on the volume of new data likely to be available at that frequency. Gains charts and ROC curves are other types of visual aid that can be useful in identifying model degradation. Rerunning the model on new data at some desired frequency and measuring the changes in parameter estimates is also insightful in measuring the continued effectiveness of your model.

Model Degradation Lift Chart



CONCLUDING REMARKS

In any organization, there are hunters (those who get the business), gatherers (those who prepare data related to the business) and scavengers (those who consume and analyze the data). Sound data is the foundation of a sound analysis. Senior management relies on analytics to make decisions that are in the best interest of the company. The processing of data for new and inforce business needs constant review and oversight from those who analyze company data. The data is most meaningful to those who consume it for analysis and decision making, and they are in the best position to inform the controls around its collection and accurate recording. Building a predictive model will waste the efforts of company talent and lead to faulty decision making if modeling data is flawed. Stay cognizant of the 80–20 rule: Modeling is 80 percent data construction and 20 percent statistical model construction. Short-changing the investment in data improvement will lead to suboptimal model building and decision making by senior management. ■

Dorothy L. Andrews, ASA, MAAA, is a consulting actuary with Merlinos & Associates, Inc. She can be contacted at *dandrews@merlinosinc.com*.

ENDNOTES

1 Morris H. DeGroot, "A Conversation with George Box," Statistical Science, vol. 2, no. 3 (August 1987), 239–258.

REFERENCES

Box, George E. P. 2013. *An Accidental Statistician: The Life and Memories of George E. P. Box*. New York: John Wiley & Sons.

Chu, Robert, David Duling and Wayne Thompson. 2007. "Best practices for managing predictive models in a production environment," SAS Global Forum 2007, Paper 076-2007, 10 pp.

Derman, Emanuel. April 1996. "Model risk," Goldman Sachs Quantitative Strategies Research Notes, United Kingdom.

Larson, Anders. December 2016. "Creating a useful training data set for predictive modeling," SOA Predictive Analytics and Futurism Section Newsletter, Issue 14, pp. 32–34.

Rud, Olivia Parr. 2001. *Data Mining Cookbook: Modeling Data for Marketing, Risk, and Customer Relationship Management*. New York: John Wiley & Sons.

Predictive Modeling Techniques—A Case Study in Resolving Correlated Explanatory Variables

By Vincent J. Granieri

INTRODUCTION

In our last article, we discussed using the Cox Proportional Hazards Model in developing a predictive underwriting model that produces a mortality multiplier for each individual. This multiplier could serve as the basis for debits and/or credits as it expresses the relative risk of having a given condition vis-à-vis not having it. This paper builds upon that foundation and presents a case study in resolving issues that we sometimes encounter when explanatory, or independent, variables are not truly independent of one another.

In fact, the predictive underwriting model we developed last time did exhibit some strange characteristics regarding cardiac structure and coronary artery disease (CAD). Because of time constraints, we glossed over these situations and applied clinical judgment to our final debit model. Now we are going to revisit this issue and see if we can't improve our model and eliminate the problem.

At the risk of being repetitious, we will include some basic information about the Cox Proportional Hazards Model so that those who are less familiar with it can get up to speed without having to consult other source material.

Cox Proportional Hazards Model

The Cox Proportional Hazards Model was introduced in 1972 as a method to examine the relationship between survival (mortality) and one or more independent, or sometimes called explanatory, variables. Some advantages of the Cox model are that it can utilize many underwritings on the same life and can handle data that is right censored; that is, subjects can leave the study at any time, or the study can end before all subjects have died. The Cox model does not require knowledge of the underlying (base) survival curve, which can be advantageous.

Cox Model results are expressed as the logarithm of the hazard, so technically, the relative risk factor for each variable is obtained by raising e to the power of the log(hazard). Actuaries will recognize this as consistent with Gompertz. The relative risk factor is interpreted just as it sounds: it describes the force of mortality acting on subjects having a certain condition relative to that acting upon the reference population, who do not have that condition. A relative risk factor of two for a condition means the subject is twice as likely to die as another subject who does not have that condition.

As an aside, we utilized the "survival" package in the R statistical language to produce our survival models. It is particularly well suited for this type of analysis. Other popular statistics programs, such as SAS, also contain survival models using the Cox model.

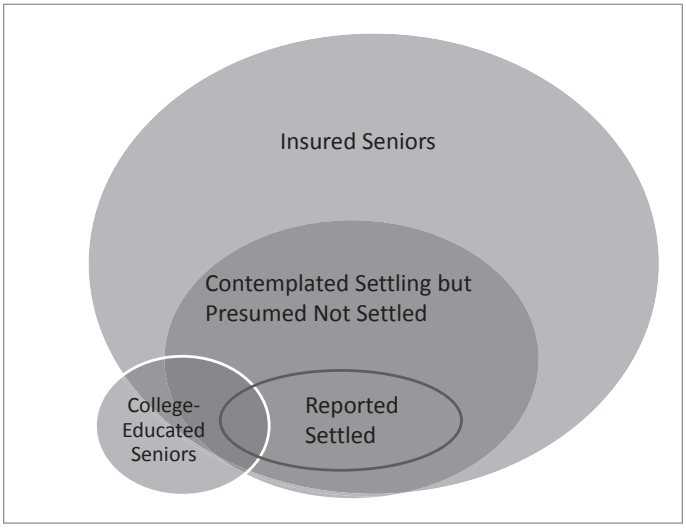
THE OBJECTIVE

Given a fully developed debit and credit model, try to resolve the confounding results observed among what seem to be similar CAD conditions.

INPUT DATA

For this exercise, we had available to us more than 200,000 underwriting events on 80,000+ unique senior lives, which took place over a 15-year period, primarily in the life settlement market. Figure 1 is a graphic description of the major subpopulations of the universe of senior lives and the populations we studied. At the highest level is the general senior population. Some of these seniors have purchased insurance, creating a subpopulation, which can be further broken out into two subpopulations: those who actually sold their policies on the secondary market, and those who contemplated such

Figure 1
Senior Populations



a sale but, for some reason, did not conclude the sale. There is also a small population of college-educated seniors, some of whom can also be associated with the other populations above. This data included demographic information such as age, gender, dates of birth and dates of death. The data also included various underwriting conditions such as BMI, smoking status and indicators for various diseases. Included were favorable conditions as well, such as family history of longevity (parents/siblings who lived beyond age 85) and good exercise tolerance.

CONSIDERATIONS WHEN EXAMINING INDEPENDENT (EXPLANATORY) VARIABLES

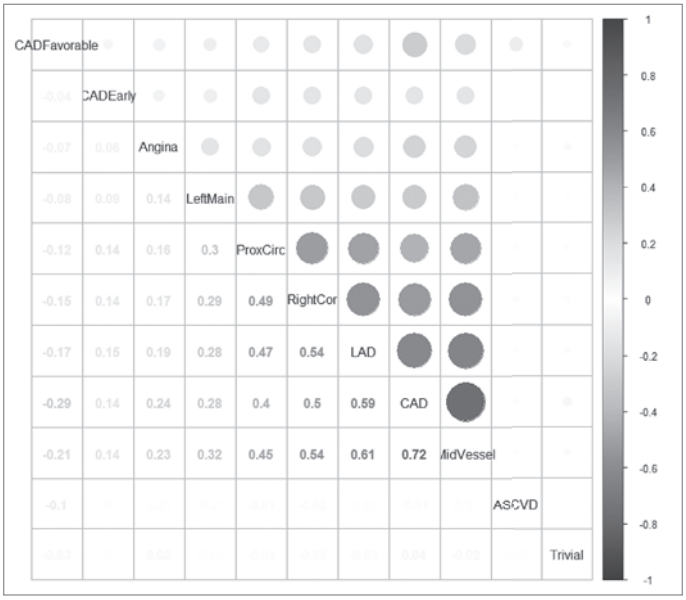
Exhibit 1 illustrates the output of the current Cox Proportional Hazards model for the CAD and Coronary Anatomy sections. Besides the name of the condition, we included a count, the number of underwritings where the subject was found with the condition, the log of the hazard, the hazard ratio (the mortality multiplier associated with the condition), upper and lower confidence intervals and *p* value. It is the hazard ratio that forms the basis for the underwriting system. Although we see many conditions whose results make perfect sense, the opposite is also true. For example, stenosis of the left anterior descending artery and one or more mid-vessel segments is seen as being protective, which is obviously wrong and problematic. Rather than use this model as is, we modified those conditions to better line up with others we felt were properly assigned debits.

When it came time to revise this model with more up-to-date data, we felt it was time to revisit this issue. We theorized that a number of these conditions were highly correlated and therefore, not truly independent. What can happen in that situation is that one variable will have overstated debits while the other may be understated. Fortunately, there is another function in R called cormat—short for correlation matrix—that quickly calculates a matrix of correlation coefficients for the variables that are input. We input the explanatory variables, and the results are seen in Exhibit 2.

Exhibit 1
Confounding Results From the Proportional Hazards Model

		Mortality Risk and CI				
Condition	**Count**	**ln(Hazard)**	**Hazard Ratio**	**95% Lower CI**	**95% Upper CI**	**P Value**
CAD Favorable — Coronary artery disease ruled out by diagnostic testing	18,006	–0.067	0.935	0.879	0.995	0.035
CAD — Atherosclerosis ASCVD calcification of large arteries	8,285	0.075	1.078	1.014	1.146	0.016
CAD — Angina current or past	2,634	–0.060	0.942	0.850	1.044	0.256
CAD — Cardiovascular disease early onset	512	0.143	1.154	0.932	1.429	0.188
CAD — Coronary artery disease	15,936	0.110	1.117	1.034	1.205	0.005
Coronary Anatomy — Stenosis of the left main	1,545	0.106	1.112	0.991	1.248	0.071
Coronary Anatomy — Stenosis of the proximal left anterior descending coronary artery	6,824	–0.008	0.992	0.912	1.079	0.857
Coronary Anatomy — Stenosis of the proximal circumflex	3,383	0.047	1.049	0.955	1.151	0.319
Coronary Anatomy — Stenosis of the proximal right coronary artery	4,987	0.031	1.032	0.946	1.126	0.482
Coronary Anatomy — Stenosis affecting one or more mid-vessel segments or secondary branches	9,228	–0.116	0.891	0.822	0.964	0.004
Coronary Anatomy Trivial	311	–0.008	0.992	0.734	1.342	0.961

Exhibit 2
Corrleation Matrix for CAD



As you can see, mid-vessel stenosis is highly correlated with a number of other blocked arteries as well as the overall CAD diagnosis. We felt that correlation coefficients higher than 0.25 were indicative of correlated explanatory variables and should be remedied somehow. But how?

With respect to the overall CAD diagnosis, we elected to eliminate it from the model. Our reasoning was that CAD was the generic term for the specific and various types of cardiac arterial stenosis. While there was high correlation in the model among the various stenosis and CAD being marked, in reality, every blocked artery condition should have also had CAD marked.

With respect to these various blockages of coronary arteries, it was becoming clear that it was quite unusual that only one such blockage would occur. We reviewed the hazard ratios that would arise if we analyzed each vessel blockage individually and discovered that a fairly narrow range of hazard ratios would ensue. We then decided to create a new independent variable, representing the number of stenosed arteries for each underwriting subject. Inserting this new variable into our model generated reasonable results, but we were not satisfied.

We felt that it was important to test whether having five arteries blocked was five times worse than having one artery blocked, for example. So we created seven new variables, each representing an additional stenosed artery from the one directly preceding it. For example, CadCAnat1 was marked when the subject had one coronary artery blocked; CadCAnat2 was marked when the subject had two coronary arteries blocked; and so on.

These new independent variables were included in the model (removing the individual variables, such as left anterior descending stenosis or right coronary artery stenosis), and the results are seen in Exhibit 3. The results indicated that having seven arteries blocked is not seven times as bad as having one artery blocked (hazard ratio of 1.74 vs. 1.19), but the results were still unsatisfactory because it was illogical that having five arteries blocked is not as bad as having four blocked (hazard ratio of 1.25 vs. 1.37), for example.

This led to another round of searching for highly correlated independent variables. Cutting to the chase, we discovered that a confirmed heart attack and bypass surgery were two more "independent variables" that were really not independent due to high correlations with the above CAD and coronary anatomy conditions. So we added those two conditions to our counts, which meant we now had nine total possible.

After rerunning the model, we saw a consistent step pattern and built new independent variables to capture the mortality risk of CAD, stenosed coronary anatomy, heart attack and bypass surgery. The final results are shown in Exhibit 4.

RESULTS

As seen in Exhibit 4, a hazard ratio of 1.35 applies to subjects with one, two, three or four blockages/ myocardial infarctions (MIs)/bypass surgeries and 1.44/1.57/1.99 for five/six/seven, respectively. The progression is logical, which was heartening. The *p* values are also miniscule, which is good. However, take good care because the tendency to find a logical explanation to justify the results of the model grows directly with the time spent building the model and cleaning data!

CONCLUSIONS

The most important conclusion is that it is a good idea to test for correlation among independent variables early on in the model building process for an underwriting system that is based on data. Given that the CAD/coronary anatomy/MI/bypass surgery portions of the model are but a small part of the total model, you can get a feel for the importance and the dominance of data preparation. We also followed this process for every other disease family in the model. Finally, this method of using counts instead of individual related conditions can produce more stable results. It is important to note that before using counts, be sure that the conditions are similar in nature and impact. Otherwise, you will find yourself averaging a high-impact variable with a low-impact variable, and your model will consistently under- or overstate the risk.

Exhibit 3
Proportional Hazards Model Results For CAD/Coraonary Anatomy Counts

		Mortality Risk and CI				
Condition	**Count**	**ln(Hazard)**	**Hazard Ratio**	**95% Lower CI**	**95% Upper CI**	**P Value**
CadCAnat1	4,020	0.1769	1.1936	1.1011	1.2938	0.00002
CadCAnat2	3,024	0.2232	1.2500	1.1339	1.3780	0.00001
CadCAnat3	2,921	0.2081	1.2313	1.1176	1.3565	0.00003
CadCAnat4	2,584	0.3133	1.3679	1.2399	1.5091	0.00000
CadCAnat5	1,972	0.2201	1.2462	1.1161	1.3915	0.00009
CadCAnat6	1,222	0.2370	1.2674	1.1154	1.4402	0.00028
CadCAnat7	463	0.5514	1.7356	1.4490	2.0790	0.00000

Exhibit 4
Final Adjustments to the CAD/Coronary Anatomy Combined Variables

		Mortality Risk and CI				
Condition	**Count**	**ln(Hazard)**	**Hazard Ratio**	**95% Lower CI**	**95% Upper CI**	**P Value**
CadAnat1to4	11,413	0.297	1.346	1.278	1.419	0.000
CadAnat5to5	1,950	0.367	1.443	1.298	1.603	0.000
CadAnat6to6	1,057	0.451	1.570	1.379	1.788	0.000
CadAnat7to7	295	0.693	1.999	1.610	2.482	0.000

SUMMARY

Regressing data to find the impact on a dependent variable of many explanatory variables is a worthwhile exercise when building an underwriting debit/credit model. However, many of the explanatory variables we access in underwriting longevity are actually correlated with one another, which confounds the models. By systematically addressing these highly correlated variables through elimination, combination and redefinition, we can improve the accuracy of the models. ■

Vincent J. Granieri, FSA, EA, MAAA, is chief executive officer at Predictive Resources LLC. In Cincinnati. He can be reached at *vgranieri@predictiveresources.com.*

Ground Assessment of Soft Skills in Actuaries

By Syed Danish Ali

ComRes undertakes a stakeholder perception audit for the Institute and Faculty of Actuaries (IFoA UK) annually, and the latest audit revealed that "people who have the most negative opinion of actuaries are actuaries themselves."[1] Andrew Brown, in a report to Institute of Actuaries, notes that "there is a universal perception that actuaries are poor communicators."[2] Elaborating on this theme, Joanne Ryan writes for the Society of Actuaries that "for many of us soft skills may not come naturally."[3]

We are aware that soft skills are important, but, somehow, they appear distant, and despite repeated attempts, we experience marginal improvements in soft skills, especially among younger actuaries.

The aim of this article is to bring this guilt out into the open and explore some possibilities as to why this might be occurring. It is only through effective understanding of the problem at hand that we can suggest some prescriptions. The second half of this article, "Building Durable Soft Skills," will explore prescriptions on how to make our soft skills robust based on understanding the problems discussed in this first part. The suggestions are based on the author's experience and observations as well as picking up pointers from reading that might help us to improve our soft skills. It is hoped that this article can help us better uncover the various hidden faces and elements that almost every actuary must confront consciously or unconsciously in our journey.

THE PROBLEM WITH SOFT SKILLS

Our soft skills usually work for us only in good times. In difficult times, we generally seem to go back to square one. This is because, normally, we are too result oriented. School fails to replicate reality because it teaches us to be successful instead of how to effectively handle failure (Nassim Nicholas Taleb—Facebook posting, December 1, 2014). We abhor making mistakes even though we know that we cannot learn anything new without making mistakes. Making and sustaining a good impression with other colleagues is foremost, and that means learning the ropes in the workplace. It is most important to be considerate of others, especially those senior to us, but we tend to overdo it and let our inner voice be drowned by their opinions.

Moreover, the explosion of information in our times has made us broad in knowing things on the surface, but perhaps inwardly shallow. Countless applications and social media incursions mean that we are constantly busy scrolling down a wall or tweeting or chatting. This has its benefits, but diverting attention to too many areas can potentially radically reduce our capacity for insight.

Socrates warned us millennia ago to "beware the barrenness of a busy life." This statement hides more than it reveals and so should be clarified. It is, of course, important to keep doing something and to be active and busy productively, but here he is warning us of "barren busyness." It has become very common to see the roles that we have to play and mold ourselves according to that (psychologists call it "mirroring").[4] Sometimes we are even proud to say that "we do not have time." It is important to spend some free time with yourself because an occupied and busy mind focusing entirely on routine will tend to not do anything to learn further or learn those skills (such as soft skills) that are not directly relevant or less important than other more immediate skills (such as quantitative skills for actuaries).

Another potential pitfall is the "fundamental attribution error."[5] This is supposing that everything bad has happened to us because of external circumstances and everything good has happened to us because of our own actions and strength. At the other extreme, this error is blaming yourself for failure even when many external influences are at play as well. Life is more complex than we usually comprehend, and in our haste to assign meaning and reasons for failure or success we tend to distort the underlying reality.

Everyday life is full of routines, deadlines and similarities. Whether it is commuting to work, attending classes or doing household chores, we experience monotony and have a gnawing feeling that we are somehow not fully alive. Weekends become a way to party and break these routines, only to become another routine itself. Life seems little more than a transaction: earn money (do study, then work) and spend money (consume brands to make life feel new and not monotonous). How many times can we say that we are doing or thinking something because we truly want it from our souls? This "barren busyness" (Socrates) causes us to lead large parts of our lives on autopilot, an automatic spiral of action and reaction, or, in other words, like an exhibition. It's just like what Plato said, "We are like people looking for something they have in their hands all the time; we're looking in all directions except at the thing we want, which is probably why we have not found it."

Sometimes, in those rare periods when we start to focus without distractions, anger is bound to be felt, but it is because we have not paid proper attention to the training of our hearts and because we associate emotions with weakness. Emotions are not

our weakness but our greatest strengths as human beings. So do any action, but from your soul, not from your mind. Feel, don't think.

BUILDING DURABLE SOFT SKILLS

The aim of this section is to elaborate on these themes so as to build soft skills within us that are not just on the surface, but hopefully deep inside our very fabric of decision making.

As actuaries, we might be focusing too much on educating our minds and not giving enough thought in comparison to education of our hearts. Aristotle has warned us here: "Education of the mind without education of the heart is no education at all." And, then, here is where Nietzsche whispers in our ears: "You must have chaos within you to give birth to a dancing star." This hits the core; we should have a healthy level of creativity, intuition and holy curiosity within ourselves as well, because this complements the scientific, intellectual and rational mind, not opposes it. The greatest of mathematical equations has to be felt first by our human emotions before it can be understood by our rational mind. We have to revive the human touch from the icy waters of calculations.

We should give due importance to the journey as well. We should accept and celebrate our failures and not just our successes. Our focus on making good impressions should be moderate and reasonable and not relying too much on our impressions, because human opinions are fickle and they take a very short time to change. We should focus on listening to our inner voice and doing what we are passionate about as well as making good impressions in the office. That means sometimes making mistakes, too, to learn something new and to progress ahead.

We need to focus more on quality rather than quantity. If we are deep enough, we will realize the interconnections between various elements and then ultimately be broad enough too.

We have to learn to differentiate between what is in our control and what is not so as to avoid the fundamental attribution error and allocate blame and praise validly.

We should not give in to the compartmentalization of knowledge where specialized and isolated pockets of knowledge are accessed without any connection to the bigger picture involved. Even if I am only a reserving actuary, I should have idea on how this connects with pricing and underwriting and risk management functions as well to have a holistic picture of the company's performance.

It is hoped that this article is able to separate the wheat from the chaff and be able to chart a course of action for us. In conclusion, to rephrase Karl Marx, we actuaries have nothing to lose but our illusions. We have a world to win! ■

Syed Danish Ali is a senior consultant at SIR consultants, a leading actuarial consultancy in the Middle East and South Asia. He can be reached at *sd.ali90@ymail.com.*

ENDNOTES

1 Cribb, Derek. July 2, 2015. "A crisis of confidence?" *The Actuary, http://www.theactuary.com/opinion/2015/07/a-crisis-of-confidence/.*

2 Brown, Andrew. May 11, 2005. "The eight habits of highly effective actuaries," Institute of Actuaries of Australia, *http://actuaries.asn.au/Library/Events/Conventions/2005/6.f-Brown_Andrew_Final%20Paper_Eighth%20habit%20of%20highly%20effective%20%E2%80%A6.pdf.*

3 Ryan, Joanne. Dec. 2014/Jan. 2015. "Hard truth about soft skills," *The Actuary* 11, no. 6.

4 Thompson, Jeff. Sept. 9, 2012. "Mimicry and mirroring can be good or bad," *Psychology Today*, *https://www.psychologytoday.com/blog/beyond-words/201209/mimicry-and-mirroring-can-be-good-or-bad.*

5 Cherry, Kendra. May 2015. "Attribution," *http://psychology.about.com/od/socialpsychology/a/attribution.htm.*

Using Python to Solve, Simplify, Differentiate and Integrate Mathematical Expressions

By Jeff Heaton

This article introduces SymPy[1], a computer algebra system (CAS) for the Python programming language. All software presented in this article is free and open source software (FOSS). When SymPy and Numpy (another FOSS package for Python) are combined with Python Jupyter notebooks, your computer becomes a sophisticated CAS. To make use of the examples presented in this article you should have Python 3.6 (or higher) installed. Additionally, the Python packages Numpy and SymPy should also be installed. Anaconda Python is the suggested Python platform for this article because of its inclusion of many packages needed for numerical computation.

At first glance, programming languages such as Python might seem very algebraic. Consider the following expression:

$$\frac{2x+3x}{2}$$

In Python, this would be written as:

```
(2*x + 3*x) / 2
```

The grouping parentheses are necessary in Python because the grouping implied by the algebraic ratio operator is not as obvious as when represented in source code. To Python (and most programming languages) this expression is simply a set of instructions that specify something to be done with x. The programming language is not concerned with simplifying the expression to $1.5x$ or other mathematical processes such as root finding, solving, differentiation or integration.

It is also important to note that because computer programs lack some of the grouping capabilities of written algebra it is always a good idea to use parentheses if you are unsure of how the programming language handles precedence. Though most programming languages follow the same rules of precedence as defined by algebra, there are exceptions. Excel is one such exception. The expression –2^2 evaluates to –4 in any programming language that I've worked with (except Microsoft Excel). The negative operator is evaluated after the power operator. However, Excel treats the negative in –2 not as an operator, but as an intrinsic part of the constant being squared. Thus, in Microsoft Excel, this expression evaluates to 4.

MATHEMATICAL NOTATION IN JUPYTER, WORD AND LaTEX

Mathematical formulas in Wikipedia are always expressed as LaTex. For example, the familiar quadratic equation in LaTex is written as follows:

```
x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}
```

In a Python Jupyter notebook, LaTex can be rendered by enclosing it in dollar signs ($):

```
$ x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} $
```

Designate this as a markdown cell (via escape m), and Jupyter renders this equation as:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

You can also right-click a Jupyter notebook LaTex rendering and export to MathML, which can be inserted into MS Word. Simply right-click an equation rendering in Jupyter and choose "Show Math As," and then "MathML Code." This will pop open a window showing an XML rendering of your equation. Copy this text into the clipboard and paste into Windows Notepad. Then recopy the text from Notepad and paste into Word. Unfortunately, the extra step of copying into Notepad overcomes some weaknesses in Words' import capabilities. I often find that I must copy/paste text through Notepad to simplify that text for Word consumption.

I've found this to be a very valuable feature of Jupyter notebooks. I often need to reference an equation from Wikipedia in Word. Trying to transcribe the equation from Wikipedia to Word's equation editor is a tedious and error-prone process. All the equations in this article were produced either by Jupyter or LaTex and imported into Word by the process just described.

LaTex is very common in the scientific community, as well as Wikipedia. Because of this, SymPy uses LaTex to display the mathematical expressions being processed.

The code presented in this article makes use of SymPy 1.0. An older version of SymPy might exist on your machine and prevent all the code in this article from working. To check the

version of SymPy installed on your machine, execute the following code from a Jupyter notebook.

```
import sympy
print(sympy.__version__)
```

This should respond with 1.0 (or later). If it does not, use the following command (from DOS/command line) to update SymPy:

```
pip install sympy --upgrade
```

ALGEBRAIC CAPABILITIES OF SYMPY

To begin using SymPy, open a Jupyter notebook and add the following lines of code as a cell:

```
from sympy import *
from IPython.display import display
from sympy.printing.mathml import mathml
from IPython.display import display, Math, Latex

x, y, z = symbols('x y z')
init_printing(use_unicode=True)
```

The **from** commands import the necessary libraries to make use of SymPy. The **symbols** definition lists the variables that will be used in algebraic expressions. For the examples provided in this article, the expressions will use the variables *x*, *y* and *z*. The **init_printing** command will allow mathematical expressions to be nicely formatted. To print mathematical equations we also define an **mprint** function, which is used to graphically render an expression:

```
def mprint(e):
  display(Math(latex(e)))
```

To demonstrate some of SymPy's capabilities, consider the following ratio of polynomials (note that ** means exponent in Python; 2**4 is 2 to the power of 4):

```
expr = (x**3 + x**2 - x - 1)/(x**2 + 2*x + 1)
```

Usually a programming language would attempt to calculate the expression, using the current value of *x*. Python would normally assign this value to the variable expr. However, since we defined *x*, *y* and *z* as SymPy symbols, something different happens. We can ask Python what type of variable *expr* is with the following command:

```
print(type(expr))
```

Python tells us that this expression is of type Add, which just happens to be the root of the expression tree. However, the point is that Python did not attempt to calculate the expression. Rather, Python stored the expression itself. We can easily turn this expression into a displayable equation with the following command:

```
mprint(expr)
```

This results in the following expression being displayed:

$$\frac{x^3 + x^2 - x - 1}{x^2 + 2x + 1}$$

This expression almost screams "simplify me," which we can easily accommodate with the following commands:

```
expr = simplify(expr)
mprint(expr)
```

This results in the following:

$$x - 1$$

Of course, this is true only if *x* is not equal to –1, or the original expression would result in a division by zero. SymPy does not check for such assumptions.

To evaluate the expression with a specific value of x, use the following code:

```
print(expr.subs(x,5))
```

This code substitutes 5 for x and results in 4.

SymPy can also solve equations. There is considerable documentation provided by SymPy to discuss equation solving. SymPy is able to solve systems of equations, differential equations, equations involving complex numbers and other options. For this section we will see how to solve a simple algebraic equation. The next section will discuss derivatives and integrals. For more details on equation solving for advanced situations, refer to Sympy's documentation on equation solving.[2]

An equation is an expression that is equal to something. In math, an expression that does not contain an equality sign is typically assumed to equal zero. In computer programming, an expression that does not contain an equality sign is assumed to evaluate to a numeric quantity that will be printed or assigned to another variable. In SymPy, equations are written using the function **Eq**. It is not possible to write the following in SymPy:

```
3*x +5 = 10
```

Though this equation is mathematically sound, it does not make sense in computer programming. In computer programming the above literally says "create an expression of 3*x+5 and assign

that expression to the constant value of 10." That is a type mismatch: an integer cannot be assigned into a expression. To create a true equation in SymPy, use the following:

```
eql = Eq(3*x+5,10)
```

This expresses the equality (and stores it in *eql*). Now that we have an equation, we can solve it:

```
z = solveset(eql,x)
display(Math(latex(z)))
```

This results in 5/3. Notice that Sympy keeps this value as a ratio, rather than creating a repeating decimal. By evaluating expressions algebraically, rather than converting everything to floating point numbers, equations can be calculated more precisely than most programming languages allow.

CALCULUS CAPABILITIES OF SYMPY

The following code demonstrates how to take the derivative of a simple formula. To test this functionality I used a question from my undergraduate calculus textbook. The derivative of sin(x) divided by x squared can be obtained by:

```
from sympy import *
x, y, z = symbols('x y z')
init_printing(use_unicode=True)
expr = diff(sin(x)/x**2, x)
mprint(expr)
```

This results in:

$$\frac{1}{x^2}\cos(x)-\frac{2}{x^3}\sin(x)$$

My textbook gave an equivalent answer, though it combined the difference into a single ratio. To test integration, we can calculate the antiderivative of the expression we just obtained:

```
expr_i = integrate(expr,x)
mprint(expr_i)
```

This takes us right back to where we started:

$$\frac{1}{x^2}\sin(x)$$

Definite integrals can be calculated as well.

OTHER APPLICATIONS

SimPy can be a very useful component of a data scientist's toolbox. At the most basic level SymPy can be used to transform a Jupyter notebook into an advanced CAS. More advanced uses allow Python code to be created to perform automated tasks that require differentiation and integration of arbitrary expressions.

I often make use of genetic programming, which can fit an actual expression to a set of training data. Genetic programming works very similarly to linear regression and neural networks, except the final model is a readable expression—the ultimate in transparency. However, genetic programs are often very unwieldy and can benefit greatly from algebraic simplification. Additionally, gradient descent can be used to optimize the coefficients of the genetic programs. By using SymPy to differentiate genetic programming–generated expressions, gradient descent can be used to optimize their coefficients.

A Jupyter notebook containing the source code presented in this article can be found at the author's github account.[3] ■

Jeff Heaton , Ph.D., is the author of the *AI for Humans* series of books and lead data scientist at Reinsurance Group of America (RGA) in Chesterfield, Mo. He can be reached at *jHeaton@rgare.com.*

ENDNOTES

1 SymPy can be obtained from *http://www.sympy.org/en/index.html*.

2 Solving SymPy equations: *http://docs.sympy.org/latest/tutorial/solvers.html*.

3 Source code can be found at *https://github.com/jeffheaton/present/blob/master/SOA/paf-sympy/sympy-soa.ipynb*.

On Building Robust Predictive Models

By Mahmoud Shehadeh

The field of predictive modeling, including parametric and nonparametric techniques, has been extensively used in the insurance industry. For example, auto insurers use predictive models to improve rating accuracy. Predictive models are employed to predict future medical costs in health insurance. In marketing departments, models help in identifying and retaining the most profitable customers. One of the main challenges analysts face in the process is to build and select stable models that can perform well in terms of prediction when applied to real future data. For model building, analysts usually rely on either a single hold-out validation or *K*-fold cross-validation. To select a final model from a pool of candidates, several methods are available. Some methods may use the *p* values of the regression coefficients or Akaike's Information Criteria (AIC) statistic, while others may rely on some accuracy performance measures such as the root mean square error (RMSE) or the area under curve (AUC). Much has been written in the statistical literature on the pros and cons of each method, and thus, they will not be discussed here. The goal of this article is to shed light on some of the issues that may not get sufficient attention and that are related to using a single hold-out validation to train the model and relying on some statistic measures to select the final model.

The article is organized as follows. We first describe the data we use in this exercise. Second, the results of an initial regression model built using a single hold-out validation are presented. Then a sampling algorithm and its results are given. In addition, we provide some well-known theoretical results to explain the empirical results. Finally, we conclude the article with some remarks.

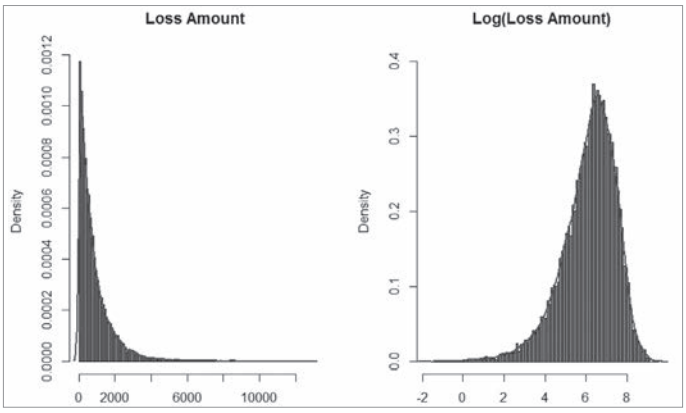
DATA

In this article we use the Pricing Game data that was published online by Dr. Arthur Charpentier. Two data sets related to auto insurance were published: Training and Pricing. The Training data set contains 100,021 records corresponding to 100,000 unique policies. The coverage exposure for the policies is for 2009 and 2010. The Pricing data set contains 36,311 records for claims that occurred in 2011. Both data sets include the same variables. The data contain the following fields: policy number, underwriting year, gender (male or female), car type (A, B, C, D E, or F), category (large, medium or small), group, occupation (employed or unemployed), driver's age, no-claim discount (a discount in the premium if no claim is made during a specific period of time), insurance contract length, car value, material cover indicator, driver's home subregion, driver's home region, population density, exposures (in days), number and total size of third party material claims, and number and total size of third-party bodily injury claims. The data sets and more details are available online at *http://freakonometrics.hypotheses.org*.

The focus in this exercise will be on the positive portion of the third-party material total cost in the Training set (about 12 percent of the data). Figure 1 shows the continuous part of loss distribution on a linear scale (left-hand panel) and on a logarithmic scale (right-hand panel) computed using histogram and kernel density. The loss distribution is positive and appears to be right-skewed. Thus, gamma regression is used to model the data. Since the loss distribution is positive and skewed to the right, we have a number of candidate models that can be considered (note that normal distribution is symmetric and will not be considered). For these data we compared the goodness of fit using graphical and numerical methods of both gamma and Weibull models, and the results came out comparable. In addition, gamma regression is often used in the industry to model the claim size.

To train an initial model, first we split the data into a training set (75 percent) and test set (25 percent), then we built a gamma regression model. The response variable (the *x* axis of the loss amount graph) is the total amount of loss divided by the number of claims (i.e., average claim amount), with the reciprocal of the number of claims as the weight. We used the logarithmic link function in fitting the model. Levels "Female," "F," "Small" and "Unemployed" are the reference groups for the gender, type, category and employment variables, respectively. Table 1 shows

Figure 1
Histograms and densities of the loss amount on a linear scale (left) and a logarithmic sale (right).



the summary of the regression model. Note that the following work is done in R statistical software (R Development Core Team 2016).

The *p* values for the model in Table 1 indicate that all the variables are highly statistically significant, except for the policy duration, value, the D and E levels in the type variable, and the Large and Medium levels in the category variable. In addition, the AIC for the model is obtained. As an accuracy measure of the model, the RMSE is computed after applying the model on the test set (the remaining 25 percent of the data).

METHOD

Next, we conducted a simulation study to investigate the variations in the *p* values, AIC and RMSE.

In this setup, suppose that we have n independent and identically distributed points (y_i, x_i), where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is the vector of predictors, and y_i is the response variable. The process is summarized in Algorithm 1.

Algorithm 1: The simulation study process.

For each $j = 1, 2, \ldots, N$, do the following:

1. Split the data (y_i, x_i) at random into two mutually exclusive subsamples $D^j_{Training}$ (75 percent) and D^j_{test} (25 percent).
2. Fit a gamma regression model $\mathcal{M}_j$ on $D^j_{Training}$, compute its AIC and collect the *p* value for each predictor.
3. Using the model $\mathcal{M}_j$ in Step 2 predict the average loss amount for each observation in D^j_{test} and compute the model's RMSE, that is,

$$RMSE_{\mathcal{M}_j} = \left[\sum_{i=1}^{n} \left(y_i^j - \hat{y}_i^j\right)^2 / n_{D^j_{test}}\right]^{1/2}.$$

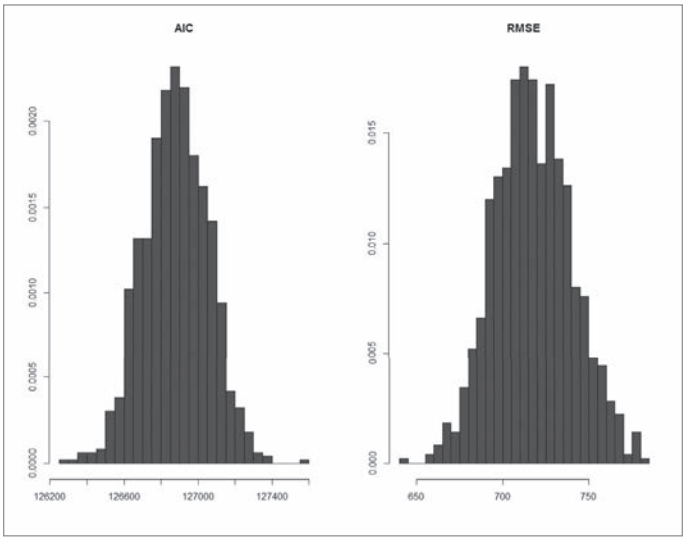
Table 1
A summary of the fitted Gamma regression model on the training set (AIC = 127,400; RMSE = 681.8)

Intercept	−85.69	41.2	−2.08	0.037597
CalYear	0.04603	0.0205	2.245	0.0248
Age	−0.0079	0.000922	−8.562	< 2-16
Bonus	−0.00175	0.000181	−9.696	< 2-16
Poldur	−0.00156	0.002269	−0.686	0.492517
Value	-1.7E-06	9.56E-07	−1.758	0.078702
Adind	−0.1221	0.02132	−5.725	1.06e-08
Density	0.000815	0.000123	6.624	3.69e-11
Gender_M	0.05702	0.02185	2.609	0.009099
Type.A	0.2806	0.04488	6.252	4.22e-10
Type.B	0.2722	0.04564	5.965	2.54e-09
Type.C	0.1647	0.04836	3.405	0.000665
Type.D	0.02476	0.04576	0.541	0.588445
Type.E	0.000287	0.04956	0.006	0.995379
Category.Large	0.004168	0.02703	0.154	0.877441
Category.Medium	0.01544	0.02469	0.625	0.531798
Occupation.Employed	−0.2418	0.02876	−8.407	< 2-16
Occupation.Housewife	−0.3775	0.03162	−11.939	< 2-16
Occupation.Retired	0.418	0.06138	6.81	1.04e-11
Occupation.Selfemployed	−0.133	0.03401	−3.912	9.23e-05

Table 2
Summary statistics of the empirical AIC and RMSE

AIC	126,300	126,800	126,900	126,900	127,000	127,600
RMSE	643.1	702	716.8	717.7	732.6	781.6

Figure 2
The histograms of the sampling distributions of AIC and RMSE.



Note that in Algorithm 1, N represents the total number of iterations, $D^j_{Training}$ and D^j_{test} are the training and test sets for the jth iteration, respectively, $\mathcal{M}_j$ is the fitted model in the jth iteration, y^j_i and $\hat{y}^j_i$ are the ith actual and predicted values of the response variable in the D^j_{test}, respectively, and $n_{D^j_{test}}$ is the number of observations in the D^j_{test}.

RESULTS

For implementation, we started the process by setting N to 1,000 iterations and apply Algorithm 1 to the training set. As a result, 1,000 gamma regression models were fitted. In Table 2 and Figure 2 we present the summary statistics and the histograms of the empirical sampling distributions of both AIC and RMSE, respectively.

Table 3 shows percentage of times that each coefficient came out significant (i.e., p value less than 0.05).

Comparing some of the results of a single model presented in Tables 1 and the simulation study in Tables 2 and 3, we note that CalYear is significant 19 percent of the time, Poldur 0 percent of the time and Category.Medium only 5 percent of the time.

Table 3
The percentage of times that each variable is significant (p value < 0.05)

Intercept	11%
CalYear	19%
Age	100%
Bonus	100%
Poldur	0%
Value	40%
Adind	100%
Density	100%
Gender_M	100%
Type.A	100%
Type.B	100%
Type.C	98%
Type.D	0%
Type.E	0%
Category.Large	0%
Category.Medium	5%
Occupation.Employed	100%
Occupation.Housewife	100%
Occupation.Retired	100%
Occupation.Self-employed	100%

Furthermore, the AIC value for the single model is 127,400, while the range of the generated AIC values from the simulation study is between 126,300 and 127,600. Similarly, the RMSE of the single model is equal to 681.8 compared to the range of 643.1 and 781.6 using the simulation.

The variability in p values can be explained using the following well-known theorem (see McCullagh and Nelder 1989) and its consequence results. Furthermore, it would be an interesting task to construct similar theoretical results for AIC and RMSE to show their asymptotic distributions, but this is out of the scope of this article.

Theorem: The sampling distributions of GLM regression coefficients that are found via the maximum-likelihood estimation with a canonical link are asymptotically multivariate normal with mean vector β and variance-covariance matrix $\phi\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}$, that is,

$$\hat{\beta} \sim \mathrm{N}\left(\beta, \phi\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\right),$$

where $\boldsymbol{X}$ is the model matrix, $\boldsymbol{W}$ is a diagonal matrix of weights and ϕ is the dispersion parameter.

In addition, the following result is a consequence of Theorem 1:

$$\frac{\hat{\beta}_j - \beta_j}{S.E.} \sim Z,$$

where $S.E.$ is $\sqrt{\phi\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}_{jj}}$ and $Z \sim \mathrm{N}(0,1)$.

Thus, we can use the z-test to test the significance of GLM regression coefficients, that is, testing $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$ with

$$z_j = \frac{\hat{\beta}_j}{\sqrt{\phi\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}_{jj}}},$$

and then computing the corresponding p values.

CONCLUDING REMARKS

To summarize, in this article we compared the results (namely, p values, AIC and RMSE) of a regression model trained using a single hold-out validation method and the results of 1,000 models trained using a simulation study. We found that these statistical results could vary from one sample to another. Thus, relying on results generated via a single hold-out validation without further investigation may produce a misleading decision. In addition, another parameter that can be used to study the variation is the portion of the training set. In this article we set it to 75 percent, but different numbers will produce different results, and it is worth investigating. ■

Mahmoud Shehadeh is a data scientist at RGA Reinsurance Company in Chesterfield, Mo. He can be reached at *Mahmoud.Shehadeh@rgare.com*.

ACKNOWLEDGMENT

We would like to thank Professor Arthur Charpentier for allowing us to use the *Pricing Game* data for this article.

REFERENCES

McCullagh, P. and J. A. Nelder. 1989. *Generalized Linear Models.* London: Chapman and Hall.

R Core Team. 2016. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. *https://www.R-project.org/*

Speculative Fiction Contest and the Predictive Analytics and Futurism Section Award

By Ben Wolzenski

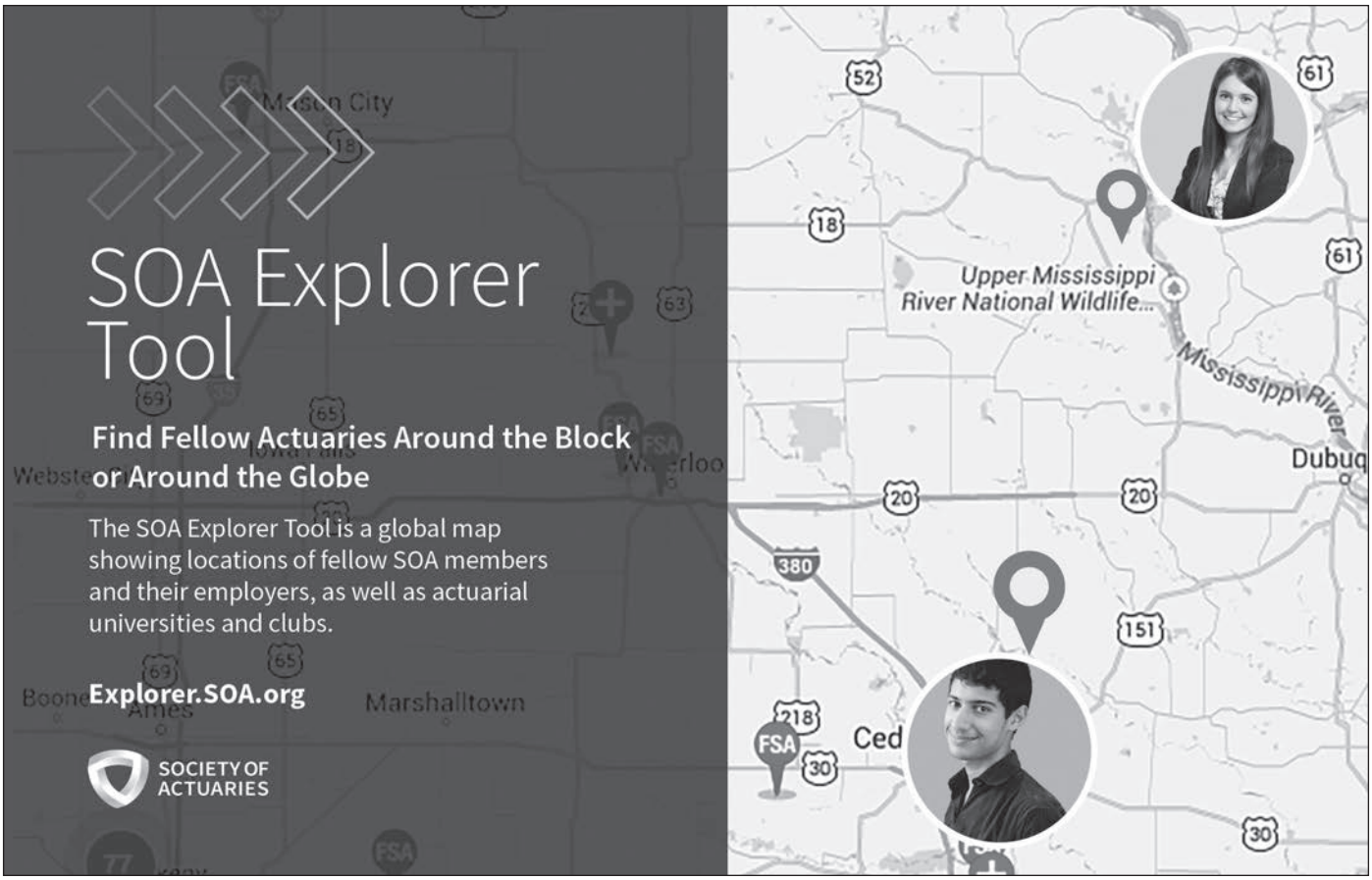
In this year's Speculative Fiction Contest there were 23 entries: short stories with a maximum length of 6,000 words. In addition to the overall award by the Society of Actuaries, the Predictive Analytics and Futurism Section (PAF) awards a prize for "the best use of futurism methods." The entries this year made enjoyable reading for eight reader-evaluators: three members or friends of the PAF Section and five volunteers who made their interest known through the new Society of Actuaries volunteer database (thanks, volunteers!).

Eighteen of the 23 stories mentioned actuaries, which was perhaps less surprising than that five of the 23 did not. Most stories were built around subjects familiar to actuaries: fourteen involved artificial intelligence, predictive analytics or models generally; a dozen more involved insurance. Other subjects used more than once were longevity, space travel, time travel and genetics.

Congratulations to the winner of the PAF award, John A. Major, ASA, MAAA, for his short story "2064: A Calculated Risk." Thanks, too, to the other 22 authors who submitted entries!

You can access all the entries on the SOA website. Use the search tool to look for "12th speculative fiction stories" and then choose the "2017" date in the results. At the time this article was written, that yielded just the 23 individual entries to the 2017 contest. ■

Ben Wolzenski, FSA, MAAA, is managing member at Actuarial Innovations, LLC in St. Louis, Mo. He can be reached at *bwolzenski@rgare.com*.



Data Visualization for Model Controls

By Bob Crompton

This article first appeared in the March 2017 issue of The Financial Reporter. It is reprinted here with permission.

One of the critical components of model risk management is effective model controls. The Committee of Sponsoring Organizations of the Treadway Commission (COSO) defines a control as follows:

"Internal control is broadly defined as a process, effected by an entity's board of directors, management and other personnel, designed to provide reasonable assurance regarding the achievement of objectives relating to operations, reporting and compliance."[1]

Examples of controls commonly used in model risk management include the following:

- Formalized approvals for model changes and updates
- Reconciliation of data
- Review and sign-off of model results
- Trending
- Ratios
- Roll-forward of accounts

Although actuaries are familiar with these types of controls, as a profession we have spent significantly more time thinking about constructing models than controlling them. Controls for actuarial models are currently full of "low hanging fruit"—that is, items that can quickly and easily be improved for a significant benefit to model risk management. One way in which we can harvest this fruit is by adding visualization to the controls we currently use.[2]

THE PROBLEM WITH CONTROLS

Many controls provide extensive numeric results from a model. These numeric results contain the potential for effective controls, but this potential is not always realized. Many controls fail to distinguish exceptions from anticipated results. They give no indication of the bounds of reasonableness and fail to provide the reviewer with indicators of where the model might be out of control.

They rely on the reviewer to make judgments regarding which items are exceptions and which are normal. Actuarial judgment is a fine thing, but it is not uniformly distributed throughout the profession. The model reviewer may not have developed sufficient actuarial judgment, or the reviewer might not be an actuary.

Furthermore, controls are often formatted in such a way that it is difficult to read and interpret the data, and even more difficult to maintain sufficient focus to apply the necessary judgments. Some controls need their own controls!

To illustrate this, a specimen control is shown in Table 1 (below).

This is from a roll-forward of universal life account values in which each of the components is shown as a change from the prior period. Even though just looking at this makes my eyes start to cross, it's clear that there is a lot of good information here, but it is difficult to tell what is what.

Table 1

1001	0.006	0.012	0.019	0.093	0.067	0.115	0.009
1002	0.015	0.013	0.024	0.077	0.000	0.050	0.007
1003	0.014	0.040	0.042	0.062	0.036	0.081	0.007
1004	0.022	0.039	0.027	0.006	0.060	0.017	0.017
1005	0.013	0.012	0.038	0.016	0.004	0.093	0.006
1006	0.004	0.023	0.034	0.013	0.072	0.009	0.015
1007	0.014	0.051	0.046	0.072	0.042	0.008	0.008
1008	0.004	0.051	0.039	0.086	0.033	0.032	0.008

Can we do better than subject model reviewers to such a painful exercise?

DATA VISUALIZATION ADDRESSES THE PROBLEM

The best controls provide immediate and effective feedback on potential model exceptions. Table 2 (below, top) is based on the data in Table 1. However, it presents the data in a binary manner—green for Exception and gray for No Exception.

Usually the simpler a control is, the more effective it becomes. Compare the ease of scanning the control in Table 2 with a more nuanced control similar in format, but with a *Consumer Reports*-style ranking shown in Table 3 (below, bottom).

Although this format provides more information than the green/gray format, it underperforms as a control because it is not as easy nor as efficient to scan.

The key to making such controls effective is understanding the normal range of results as well as what typically causes outliers. The model owner will need to articulate this understanding in such a way that the quantification of the range of normal results is possible. As an example, the model owner for the roll-forward model shown above may have determined through experience that any unallocated amount of fund change greater than ±2 percent of the fund is indicative of an outlier. On the green/gray control above, any unallocated amount more than ±2 percent would show up as a green light.

Both the green/gray control and the *Consumer Reports*-style control were created in Excel, using conditional formatting.

SOME GENERAL RULES FOR VISUALIZATION IN CONTROLS

The difference in the efficiency between the two ranking controls above points us to some of the rules for data visualization controls. Since visualization is more of an art than a science, these rules are stated in general form. The practitioner must decide how these are best applied in any situation.

- Make controls as simple as possible, but as complex as necessary
  - Controls should provide only the information needed to determine the control decision
- Provide immediate indications of actuals versus expectations
- Emphasize the critical data
- Changes in output values are often more informative than either the beginning or ending values

Table 2

1001								
1002								
1003								
1004								
1005								
1006								
1007								
1008								

Table 3

1001								
1002								
1003								
1004								
1005								

- Orient the data in the most user-friendly way
- Color draws the eye quicker than black and white
- Use a visualization style suitable to the purpose—for example:
  - Line graphs work well for trends
  - Bar charts work well for rankings
  - Maps work well for geographical data

The goal is to make the data visualization work as a process con-

> Make controls as simple as possible, but as complex as necessary.

trol chart—a tool that quickly tells the model reviewer whether results are outside of the boundaries of reasonableness.

WHEN REASONABLENESS BOUNDS CANNOT BE EASILY ARTICULATED

In some instances, the modeler will have difficulty articulating what the bounds of reasonableness are for modeled items. This may be due to the multifactorial nature of the item, or it may be due to the nonlinearity of the item. It could be due to both the multifactorial nature and nonlinearity.

Whatever the reason for the difficulty, the modeler will usually only have a rough sense of how modeled values will emerge from the model.

A typical example of this sort of model item is the reserve per $1,000 of in force that is often used as a control for valuation models. There are various forces that affect the reserve/$1,000 for any particular valuation cell, including:

- Number of policies in the cell
- Amount of in force in the cell
- Type of benefit
- Premium paying pattern

So this is definitely a multifactorial item. In addition, the slope of reserves is usually nonlinear, adding to the difficulties in determining the bounds of reasonableness.

Not only is it hard for the model owner, it is also difficult for the auditor. The PCAOB has come down very hard on auditors for not giving sufficient scrutiny to this sort of control, and for not documenting their analysis of the effectiveness of the control. The following quote from Helen Munter, director of the Division of Registration and Inspections of the PCAOB emphasizes this point:

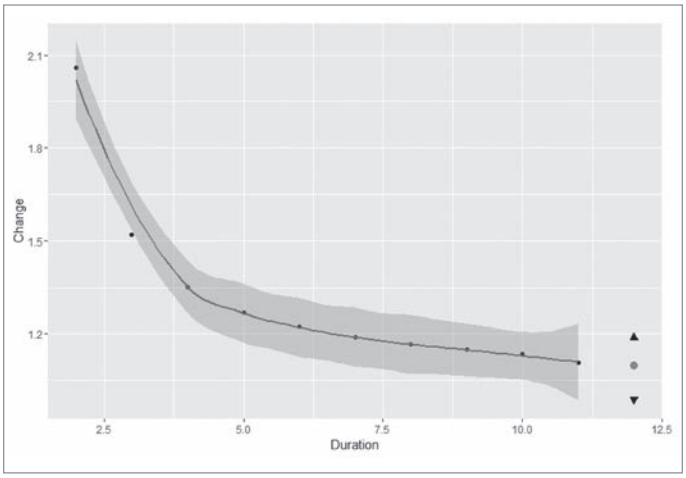
> Over the last few years, the audit of internal control has topped the list of deficiencies in the audit work we have reviewed.[3]

When the required articulation is not possible, it is still possible to develop visualizations for the bounds of reasonableness. We require a general fitting method combined with predictions of the model item in question. Figure 1 shows one such approach.

In Figure 1, the dots in and around the shaded area are historical actual reserve change ratios. The line inside the shaded area is the curve fitted to the data. The shaded area is the fitted curve plus/minus one standard error.

This approach used loess regression (a nonlinear approach in which a series of polynomials is fitted to the data) for the first 11 policy durations, and a prediction interval for the 12th duration is given as the point estimate ± one standard error. These bounds of reasonableness are shown as triangles, while the actual result is shown as a circle. In this example, we see that the actual result falls comfortably within the bounds of reasonableness.

Figure 1



It is possible to programmatically chart a series of such reserve progressions. It is also possible to export the results into a Red/Green indicator type spreadsheet in addition to (or in place of) charting the results as in Figure 1.

REVIEW AND SIGN-OFF CONTROLS

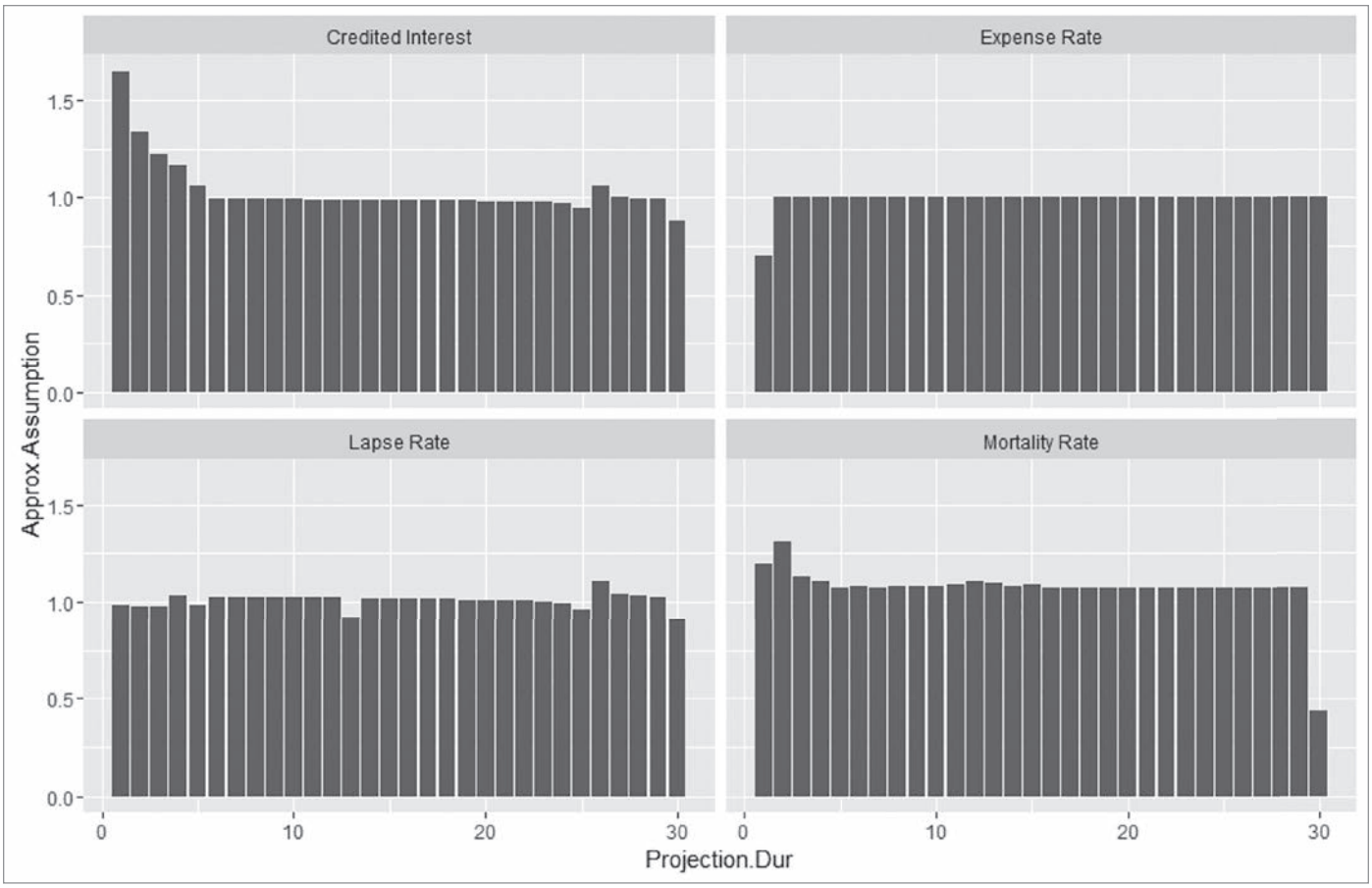
Review and sign-off controls are subject to several difficulties. Sometimes the sign-off form merely states that the model has been reviewed for reasonableness. (Occasionally there will be sign-off forms that merely assert that a review has been performed, but most companies seem to have realized the true value of this assertion.)

A simple assertion of reasonableness is troublesome from several aspects. The first is that it might not be clear precisely what model output has been scrutinized for reasonableness. It is possible that several items could be effectively reviewed for reasonableness, yet a critical model output might not be scrutinized. Such an oversight could easily go undetected until there is a material model problem.

> Data visualization is limited mostly by our imaginations rather than our software capabilities.

Another troubling aspect of such a review is that there is no definition of what constitutes reasonableness or of where the boundaries of reasonableness lie. If the reviewer has different judgments on reasonableness compared to the model designer or the model owner, then we should expect either false model exceptions or missed model exceptions.

Figure 2



A final difficulty with such a simple assertion is that if it is time sensitive, the depth and extent of the review could be subject to variability.

In order for a sign-off control to work uniformly, there needs to be a structure provided in which the review takes place. Often what is wanted in a reasonableness review is a review of the directional changes in model output compared to the directional changes in model assumptions. One way to address this is to put a visualization of the ratio of stated directional changes versus approximated directional changes into a quickly and easily assimilated visualization. The example in Figure 2 (pg. 21) shows the ratio of the documented assumption versus the approximation of the assumption calculated from model output.

In this visualization, the significant drivers of model output are shown together in order to ease the reviewer's job. The reviewer would need to decide if the early-duration and late-duration variations are true exceptions or if they are artifacts of the approximation methodology.

Another item that may be of interest is model composition such as in force by issue age or underwriting category. One way to quickly display such information is in an ordered bar chart such as Figure 3 (below).

For model control visualization, we can put together a historical series of charts for some selected number of past model cycles in order to provide an additional dimension to the visualization.

WHEN VISUALIZATIONS GO WRONG

One of the more popular forms of visualization found on many websites is the "mosaic plot." A mosaic plot display of the information in the In Force Composition from above is shown as an example.

Figure 3

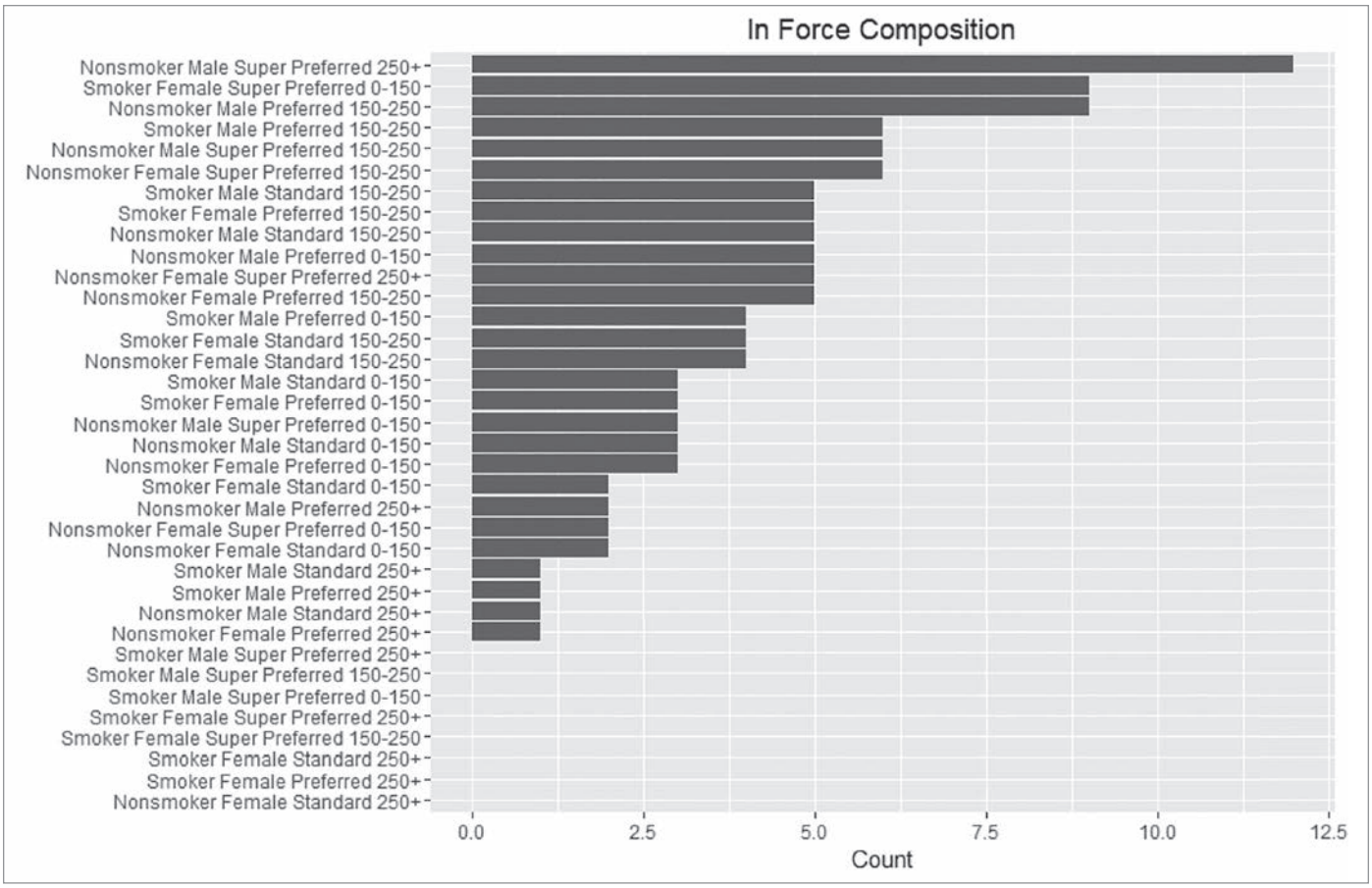
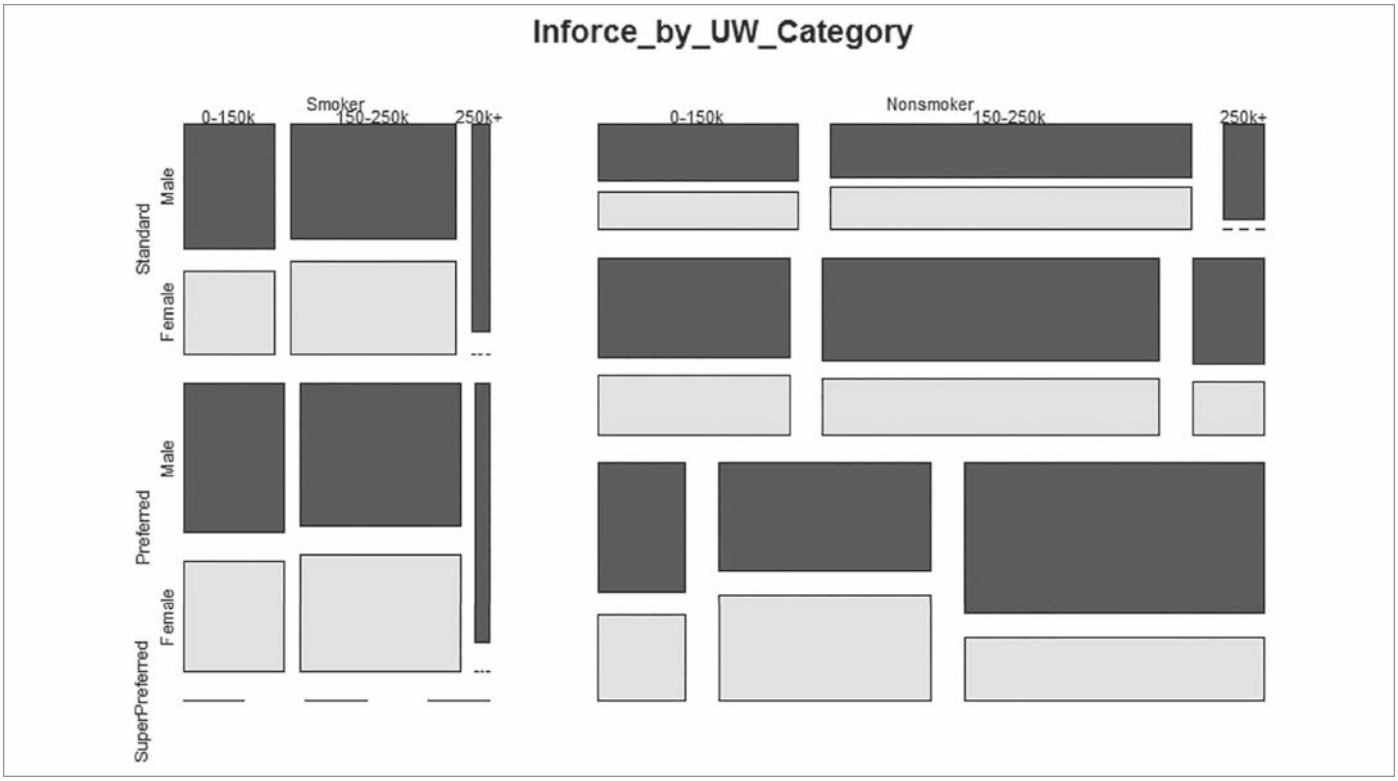


Figure 4



Mosaic plots are interesting and fun to look at, but they don't work as control visualizations. A brief scan of the visualization in Figure 4 shows that it is difficult to make quantitative comparisons between different segments, or even to quickly determine the largest segments of in force.

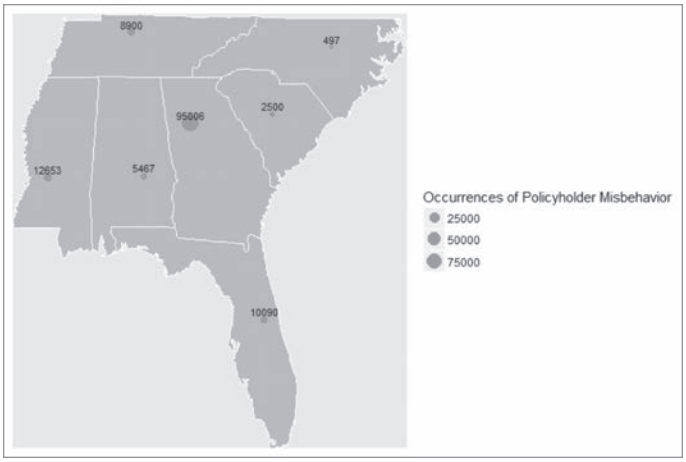
Cells with similar areas sometimes have markedly different dimensions—this issue is so profoundly non-intuitive that it is difficult to conceive of any situation in which a mosaic plot would make an effective visualization for a model control.

Just because we can create a visualization doesn't mean that we should create a visualization.

GEOGRAPHICAL DATA

Whenever a model creates output with a geographical component, maps become an option as a control item. A well designed map provides more information per pixel than almost any other visualization. In the hypothetical example given in Figure 5, I have shown a projection by state of the number of policyholder misbehaviors. Policyholder misbehavior is any activity that results in adverse results for the insurer. The visualization provides a quick relative comparison as well as providing precise information regarding the number of projected occurrences.

Figure 5



CONTROLS FOR WHEN THERE ARE NO BRIGHT LINES

In situations where we are not circumscribed by prescribed methodologies or assumptions, we might be interested in a "better/worse" comparison rather than "reasonable/unreasonable" comparison. Appraisal models and planning/budgeting models might fall into this category.

For such better/worse comparisons, a heat map might provide quick information regarding the relative performance of model output compared to some standard of expectation. Heat maps highlight worse results with "uncomfortable" colors while highlighting better results with "comfortable" colors.

In Table 4, a heat map is used to show how model output compares to projected historical trends.

Table 4

Item	2017	2018	2019
Premiums	3.3%	6.0%	9.3%
Death Benefits	-5.1%	-7.2%	-8.0%
Lapse Benefits	-6.2%	-8.7%	-11.7%
Expenses	20.8%	37.5%	61.5%

- Differs from trend by ± 5%
- Differs from trend by ≥ 5%, < 10% absolute change
- Differs from trend by ≥ 10% absolute change

This heat map was created in Excel, where conditional formatting makes such visualization easy.

There is an interesting issue hidden in the implicitness of numbers used to construct the heat map. The standard for Better and Worse was a linear trend line. Why did I choose a linear trend? Mainly for illustrative purposes. In real life, some nonlinear form of trending might be more appropriate, and might be a better reflection of what is reasonable.

In all of these examples, experience and a firm grasp on reality are important in setting the bounds of reasonableness. As Salvador Dali, the great surrealist, might have said:

> One person's reasonableness is another person's melting watch.

CONCLUSION

Actuarial model controls are ripe for improvement. One way to greatly enhance the effectiveness of many controls is to include some form of visualization. Visualization can be done with spreadsheets, with R or with some form of commercial data visualization package. Data visualization is limited mostly by our imaginations rather than our software capabilities. Many other forms of visualization are possible and will no doubt come into practice as actuaries focus more on controls. ■

Bob Crompton, FSA, MAAA, is a vice president of Actuarial Resources Corporation of Georgia, located in Alpharetta, Ga. He can be reached at *bob.crompton@arcga.com*.

ENDNOTES

1 From the document "Internal Control—Integrated Framework" on COSO's website at *http://www.coso.org/documents/990025P_Executive_Summary_final_may20_e.pdf.*

2 The visualizations shown in this article were created using R software, except where noted differently.

3 *https://pcaobus.org/News/Speech/Pages/Munter-Audits-Internal-Control-IAG-09092015.aspx*

Using Predictive Modeling to Risk-Adjust Primary Care Panel Sizes

By Anders Larson

Most health actuaries are familiar with the concept of risk adjustment. Some of the most well-known uses in the health insurance industry include using risk scores to help determine payment rates for Medicare Advantage plans, transferring funds between commercial plans on the ACA exchanges, and adjusting capitation rates for managed Medicaid plans. It is also common for insurers, self-funded employers and providers to use risk scores to account for differences in morbidity between different populations for a variety of other purposes.

In many cases, risk adjustment models use diagnosis codes and other information from claim and enrollment data to produce risk scores that predict total costs, or at least predict a significant portion of total costs (for instance, medical or pharmacy costs only). However, risk adjustment does not necessarily need to be defined so narrowly. Depending on the intended purpose, "risk scores" are not required to be based strictly on diagnosis code information, and they are not required to predict total costs. For purposes of this article, we will define a risk score as a quantitative model that makes a prediction about health care utilization or expenditures. For some applications, it may be important for the model to make the predictions based on patient characteristics that are not controlled by the parties at financial risk (often a payer). One example of a risk score predicting something other than claims costs is the Framingham Risk Score, which predicts the 10-year cardiovascular risk of an individual, based on age, gender, cholesterol levels, smoking status and blood pressure.

This article discusses another nontraditional use of risk adjustment that incorporates modern predictive modeling techniques: risk-adjusting primary care panel sizes. We will describe the business problem, available data sources and challenges specific to this assignment, as well as the statistical techniques used to develop the risk scores.

THE BUSINESS PROBLEM

Provider reimbursement has shifted from a largely fee-for-service model in recent years to include value-based contracts between payers and providers. This paradigm shift has also extended to compensation models within provider organizations. For instance, primary care physicians are often compensated based on the number and intensity of services they provide, regardless of the number of unique patients they serve. In that case, seeing a single patient 10 times generates roughly the same income as seeing 10 patients once each. This system can create an incentive for physicians to bring patients in for more services than are necessary. In turn, this also limits the physician's ability to open the practice to additional new patients.

If the goal of the primary care organization is to provide appropriate care to the maximum number of patients, the organization needs a way to determine the appropriate number of patients for each physician (panel size). Of course, all physicians do not serve the same type of patients, and it would be unrealistic to expect all physicians to have the same panel size, even if they work the same number of hours. So what **is** the appropriate panel size for each physician?

The way we approached this problem was to develop a customized model to predict the number of primary care visits each patient should require over the next six months. The prediction was based on a wide variety of patient characteristics, including demographic information, clinical conditions and historical utilization of certain health care services, such as emergency room visits and inpatient admissions. The model did **not** base the predictions on each patient's historical office visit utilization or which physician they were assigned to. If these features were included, physicians who had been seeing their patients too frequently in the past would have their patients receive predictions that were higher than similar patients who saw other physicians. It is true that excluding these features reduced the predictive power of our model, but this was necessary to achieve the specific business objectives.

Ultimately, the predicted office visits were converted to office visit time for the physician's current patient panel, and the

predicted office visit time was compared to the physician's scheduled working hours over the next six months to determine if the physician had capacity to add new patients. The predicted office visit time for each patient could also be used to help facilitate more useful comparisons of "risk-adjusted panel sizes" between physicians. For instance, if an average patient required 30 minutes of office visit time per six months and a physician's current panel of patients was estimated to require 30,000 minutes of office visit time over the next six months, we would say this physician had (30,000 / 30) = 1,000 risk-adjusted patients. The number of risk-adjusted patients divided by the number of actual patients represented the panel's average risk score.

CHALLENGES WITH AVAILABLE DATA SOURCES

Providers, including primary care physicians, typically see patients who are insured by a variety of payers (and some patients who are uninsured). Therefore, using paid claims data from insurers, which actuaries most commonly rely on for analysis, was not a viable data source in this case. Instead, we used billing data from the provider organization, which included some of the same fields as paid claims data: service dates, provider ID, ICD diagnosis codes, CPT codes, place of service and billed charges (but not plan paid or allowed amounts). Our analysis incorporated billing data from three years for more than 200,000 patients, which allowed us to develop a very robust model.

One challenge with this data source was that there was no concept of "enrollment," which would typically exist with paid claims data. This presented two problems:

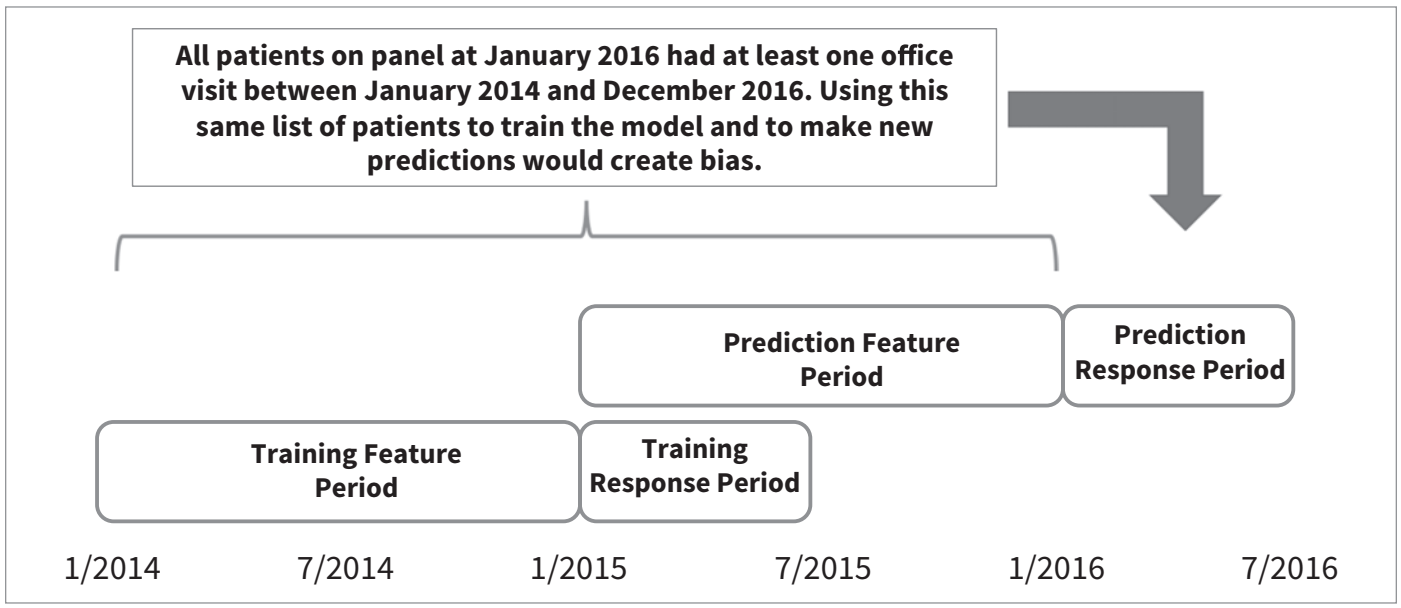
1. The data included all services that had occurred with the provider organization over a specific period, regardless of whether the patient was seeing a primary care physician within the provider organization. Selecting which patients and services should be included in our analysis was critical.

2. If a patient did not have any services over a period, there was no clear way to determine whether a patient was "eligible" for services and simply did not have any, or whether the patient was not really "eligible" to receive services. For instance, a patient who moved to the area in January 2016 would not have received any services in our data in 2015, but it would not be accurate to say this patient was not receiving any services at all in 2015.

To address the first problem, we limited our analysis to data for two sets of patients: all patients on the current primary care panel[1] and all patients on the primary care panel as of a specific date in the past. The data for the first set of patients was needed to determine the characteristics of the current panel of patients, for whom we would be making predictions. The data for the second set of patients, however, was equally critical: this would be the data used to train and calibrate our predictive model.

In predictive modeling, the data used to train the model should be a reasonable representation of the data used to make predictions. Figure 1 shows the time periods used in our analysis. In our case, we trained the model by looking at the relationship between patient characteristics in 2014 (training feature period) and utilization in the first half of 2015 (training response period).

Figure 1
Training and Prediction Periods



For this provider organization, the patients were included in the primary care panel only if they had seen this group of primary care physicians in the past two years. If we used data for the January 2016 panel of patients to train the model, we would necessarily exclude anyone who dropped off the panel in the past year. Certainly some patients on the current panel would drop off in the future, and these types of patients needed to be represented in the training data set.

It was more difficult to address the second problem (interpreting periods of inactivity). One option was to consider a person "eligible" for all months after their first observed service. Although this approach was reasonable, we were concerned that utilization rates would be distorted for patients whose first visit occurred relatively recently due to the low amount of "eligibility." In the end, we opted not to estimate periods of eligibility at all. Patients on the primary care panel were not differentiated based on the date of their first service, although we did include a binary variable indicating whether the patient was appearing on the primary care panel for the first time (i.e., their first service had been in the most recent month, since the primary care panel was updated monthly). These patients were clearly very new and might require extra office visit time in the next few months.

SELECTING THE PREDICTIVE ALGORITHM

Many popular risk-scoring algorithms are based on some type of linear model. For instance, the CMS-HCC model used in the Medicare program assigns a coefficient to each of approximately 80 conditions, and each patient's risk score can be calculated by summing the coefficients for the conditions observed for that patient, plus an additional value related to the person's age, gender and enrollment category. Although there are some exceptions, the model generally does not account for interactions between conditions or differences in how a condition might impact patients differently at different ages. For example, the value of congestive heart failure is the same for a 90-year-old male and a 65-year-old female.

While linear models have the advantage of being relatively easy to understand and interpret, they are often outperformed by other modern machine learning algorithms. In many cases, industry standards and generally accepted practices also limit the ability for many risk-scoring algorithms to use more complex models. Since this was not the case for this assignment, we were open to different approaches. We found early in our work that decision-tree-based models produced more accurate results than a generalized linear model (GLM), even when the two models used the same features. Among the reasons for this were multicollinearity between features, the large number of available features, and clear nonlinear relationships between certain variables, such as age and office visits.

Given the computing power available today, it is rare to use a single decision tree algorithm in modern predictive modeling. Instead, predictions are often derived from large numbers of decision trees, referred to as ensembles. The two most common ensemble techniques are boosting and bagging. In our case, we opted for a boosted decision tree model known as a gradient boosting machine (GBM). Although using a bagging algorithm such as a random forest would have likely produced satisfactory results, the GBM had the advantage of being able to properly model a conditionally Poisson response variable. In our case, we were interested in predicting a count of office visits for each patient, which was commonly zero, one or two.

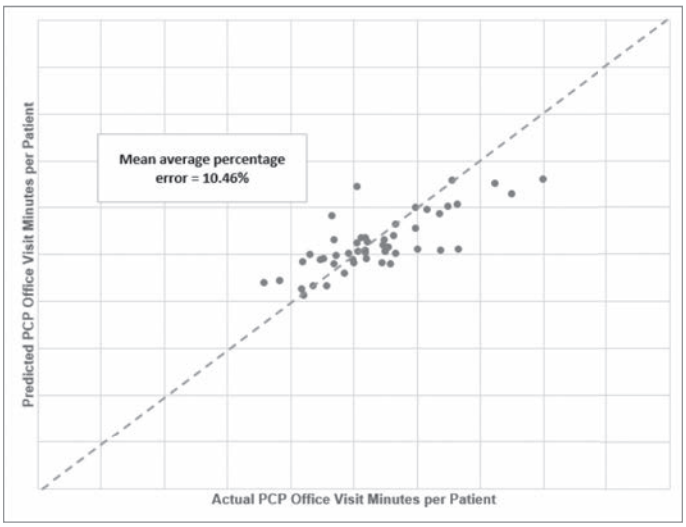
To avoid overfitting, we used a technique known as cross-validation. Cross validation involves training the model on a portion of the training data and testing the fit of the model on the remaining training data. This is repeated for other slices of the training data to get a realistic estimate of the model fit with different hyperparameters. In our case, we used 10-fold cross validation, meaning we split the training data into 10 cohorts to perform the cross validation.

Figures 2 and 3 show the model fit for the physicians with a credible number of assigned patients, both with the GLM and GBM models. The green dotted line indicates where "perfect" predictions should fall. Although the predictions are similar, the GLM model has more "big misses" where the predicted results were far from actuals, several of which can be seen on the far right of Figure 3.

The value of our model was not derived solely from its predictive accuracy. A "black box" model would be unlikely to get buy-in from physicians, regardless of how impressive the error metrics might be. We needed to provide some indication of what features were driving the results. Since decision-tree–based models do not have coefficients in the same way that linear models do, other techniques are needed for determining feature importance. In our case, we utilized a relative influence method that is based on how much each feature reduced the Poisson loss function. One way of interpreting this metric is that it indicates how much predictive power would be lost by removing each feature.

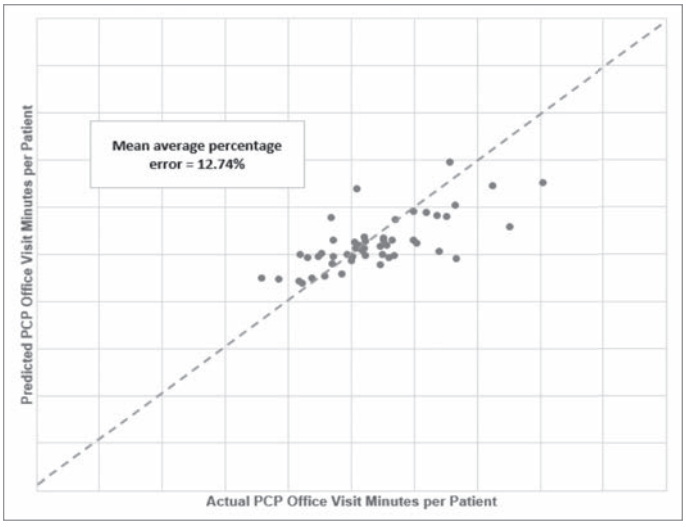
We also removed many features that appeared to have low relative influence. We found that instead of using a list of approximately 120 clinical conditions as features, we could achieve almost identical predictive accuracy by using only eight specific conditions, plus a simple count of the number of other conditions. Limiting the number of features allowed us to communicate our results more easily to physicians, who could verify whether the relationships identified by the model were intuitive.

Figure 2
GBM Model Fit



Note: Mean average percentage error was calculated including all PCPs, including those not shown in Figure 2.

Figure 3
GLM Model Fit



Note: Mean average percentage error was calculated including all PCPs, including those not shown in Figure 3.

CONCLUSION

There is no one-size-fits-all solution to risk adjustment. As the health care delivery system continues to evolve, the applications of risk adjustment are likely to evolve as well. The concept of risk adjustment can be applied to specific types of services and can be used to achieve a variety of business objectives. However, more innovative or nontraditional uses of risk adjustment sometimes require models that are customized for the particular situation. That may mean applying modern machine learning algorithms, as we did in this case, but that is not always required. If a simpler model is able to achieve a similar level of predictive accuracy, there may not be a need to use a more complex model. Even with a simpler model, however, care must be taken to calibrate the model on a data set that appropriately reflects the data that will be used to develop predictions in the future and to take steps to ensure the model is not overfitting the calibration data. In many cases, this is the most challenging and crucial part of the process.

Despite the challenges (or perhaps because of the challenges), actuaries with a combination of health care subject matter expertise and strong predictive modeling abilities are well positioned to be leaders with risk adjustment. ■

Anders Larson, FSA, MAAA, is an actuary at Milliman in Indianapolis. He can be reached at *Anders.larson@milliman.com.*

ENDNOTES

1 The primary care panel is a list of all current patients assigned to any primary care physician in the organization. This list is updated on a regular basis to add new patients and remove patients who are no longer considered current. At the time of our analysis, the "current" panel was from January 2016.

Bayesian Inference in Machine Learning

By Denis Perevalov

As the amount of data keeps growing, machine learning is drawing interest from different fields. With more data, one could find patterns and potentially use them in forecasts and recommendations. Maximum likelihood estimations (MLEs) are the most widely used machine learning methods, which is due to their speed and scalability. However, when dealing with smaller amounts of data or when data is narrow in the longitudinal direction, Bayesian analysis is arguably a better approach. Not only can it make more precise predictions, but its confidence intervals of model parameters are more interpretable.

Machine learning can be defined as the process of learning a predictive model's parameters from data. For a full specification of a problem, one has to have three ingredients: data, a predictive model hypothesis with parameters θ and a specification of the likelihood of observations, given the model and a set of predictive variables:

$$L(y|\theta, X)$$

where y is a vector of observations and X is a matrix of predictors.

The task of machine learning is the following: Given a training set of data (y, X), make the **inference** or **best estimate** of θ. In MLE, the latter is the one that yields the highest total likelihood in the training set:

$$\hat{\theta}_{best} = \text{argmax} L(y|\theta, X)$$

In the Bayesian approach, instead of a single point estimate $\hat{\theta}_{best}$, we predict a probability distribution function (PDF) of θ. We use the famous Bayes formula:

$$P(\theta|y, X) = \frac{p(\theta)L(y|\theta, X)}{\int p(\theta`)L(y|\theta`, X)d\theta`}$$

$P(\theta|y, X)$ and $p(\theta)$ are called **posterior** and **prior** distributions of θ, respectively. The integral in the denominator is a normalization constant, which is usually not important because we are interested in relative comparisons of θ.

The main feature of Bayesian analysis is that there is no optimization involved—it is simply a calculation of the posterior. However, the calculation should be performed for every single point in the space of θ. This is obviously unfeasible. Thus, we have to rely on the approximation of the posterior using samples of θ. In lower dimensions of θ, it is possible to do random sampling for the posterior estimation. In higher dimensions, one has to use more sophisticated sampling techniques. These techniques do not sample the entire θ space, but only its most likely part, and they still deliver an unbiased posterior estimation. Finally, because there is no optimization involved, there is no **overfitting** problem in the Bayesian inference.[1]

The prior distribution $p(\theta)$ is an assumption for the θ distribution before inferring it from the training data. It could be **informative** or **noninformative**. People talk about informative prior to when there is a good understanding of the θ distribution, which usually comes from an inference from some other data prior to the current study and results in a relatively narrow $p(\theta)$. Noninformative prior is used when conducting the study for the first time and when there is a very vague understanding of the θ distribution, maybe in terms of wide ranges. In that case, very wide prior distributions are used, such as normal with very high variance or uniform distribution with high width. Posterior lies somewhere within the prior distribution and is usually much narrower than the latter.

One has to be careful when choosing the prior distribution. For example, if one chooses prior to be uniform [–1,1], then posterior will always be somewhere in this interval, no matter what data suggests. The "Frequentism and Bayesianism" blog post has other good examples where a poor choice of prior may significantly bias the posterior.[2]

The main advantage of the Bayesian approach is that its result is a much richer description of the possible values of the model parameters. Apart from prior, the MLE result is only a special statistic for the Bayesian result—it is approximately its mode. MLE describes a single point in the θ space that is most likely, whereas a Bayesian result provides an entire distribution. For example, one could immediately calculate a mean, variance and skewness. In the case when one is interested in the expected value of θ, the mean is a more appropriate statistic than the mode, especially for highly skewed θ posteriors. Moreover, one could use the Bayesian θ posterior straight to infer the parameters' confidence intervals and infer possibly nonlinear correlations among individual parameters whereas MLE has to rely on variance approximations.

The disadvantage is that it usually takes much more time to fit a Bayesian model. Also, the result contains samples of the θ distribution, which may take a lot of disk space. Recent advances in the sampling algorithms, and in general having more computing power, have made it applicable to larger data sets. Currently the rule of thumb is that it is useful for data sets with at most tens of thousands of data points.

For a more detailed comparison of the Bayesian and MLE approaches please refer to the outstanding blog post "Frequentism and Bayesianism," cited above.

SIMULATION

In this article we will consider a hypothetical problem in the context of a variable annuity (VA) product and apply the Bayesian approach. Simulation and visualization are done in R, whereas Bayesian inference is done using a probabilistic language, Stan.[3] All the code, including the main Jupyter notebook, can be found on GitHub.[4] This may serve as a good-use case example for the reader.

We are going to simulate the following toy model. We will have 100 or 1,000 simulated people in our study. For each person we have 10 consecutive observations. Each observation is a binary event (i.e., 1 or 0), whether a person took a withdrawal in the given quarter or not.

The 100 or 1,000 simulated people samples will have the same random number seed, so that for the first 100 people both samples are identical. This is so that we can observe how adding more data helps in the inference of parameters.

Each person has a base withdrawal probability that can be different from other people. For example, maybe there is another parameter (income or credit score) that we do not have data for that affects the person's withdrawal probability. In our simulation, the base withdrawal probability is drawn from a normal probability density function (PDF) with a predefined mean and variance. Once drawn, it stays the same for this person. However, we allow the withdrawal probability to change with time (quarter number dependence).

To make the model more realistic, we will also allow the probability of withdrawal in a given quarter to depend on the pattern observed before that. Namely, the logit probability will have an instantaneous jump right after the first withdrawal event. This is to simulate the fact that once the first withdrawal happens, there is much higher probability that the person would withdraw in the next quarter than before that.

A one-person simulation algorithm is as follows. With predefined overall constants μ, σ, *CWD* and *Cq:*

1. Draw base logit probability from normal distribution:

$$\text{base\_logit} \sim N\left(\ , \sigma^2 \right)$$

2. Initialize withdrawal indicator $WD_{IND} = 0$
3. Loop q from 1 to 10
   - Calculate quarterly withdrawal probability (quarter count starts from 1):

$$\frac{1}{1+\exp\left(-\text{base\_logit} - C_q\left(q-1\right) - C_{WD}WD_{IND}\right)}$$

   - Draw the resulting withdrawal observation (0 or 1) for the current quarter from Bernoulli distribution:

$$WD_q \sim Bernoulli\left(p_q\right)$$

   - If WD_q == 1, then set $WD_{IND} = 1$. If the first withdrawal happens, set the indicator to 1.

Both μ and σ define the base logit withdrawal distribution, *Cq* defines withdrawal probability dependence on quarter number and *CWD* defines an instantaneous jump in the withdrawal probability after the first observed withdrawal.

We are interested in the inference of overall model constants μ, σ, *CWD* and *Cq*, as well as base logit probabilities for individual people.

EXPLORATION

In Figure 1, one can see the simulated withdrawal probabilities and withdrawal events for the first two simulated people. The first person turned out to have a much lower base withdrawal probability than the second person, by about 40 percent. We can also observe this effect in the observed withdrawal events. The second person has a higher number of withdrawals: eight versus four.

The first person's first withdrawal happens in the third quarter. For the second person, the first withdrawal is in the second quarter. Right after the first withdrawal we can observe an instantaneous jump in the simulated withdrawal probabilities for both of them, by about 6 percent in this case.

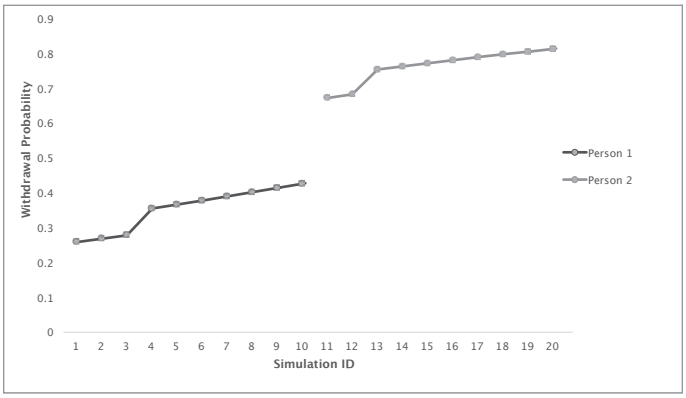
There is a roughly linear increase in probability of withdrawal with the quarter number.

BAYESIAN MODEL AND RESULTS

For the Bayesian inference we used Stan. Stan has an interface with R, the rstan package. When using rstan, one can construct data in R, launch Stan inference and get results back in R. All the code is available on GitHub. From the example, one could see that programming in Stan is fairly straightforward. All that's required is to specify the data structure, declare model parameters and specify the model—both prior and likelihood. When you pass data and the model code to Stan, it produces posterior distributions for the model parameters.

In this case we have 4+N people model parameters: four coefficients—μ, σ, *CWD* and *Cq*—and a base logit probability for each person in the training sample. Performing inference using MLE with this many parameters would be problematic because of a high chance of overfitting. However, as discussed earlier, in the Bayesian approach there is no overfitting.

Figure 1
Simulated Withdrawal Probabilities



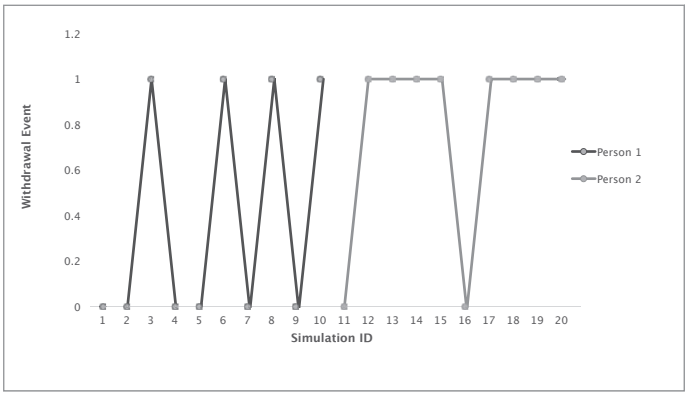


Figure 2
Inferred Distributions for μ, σ, C_{WD} and C_q

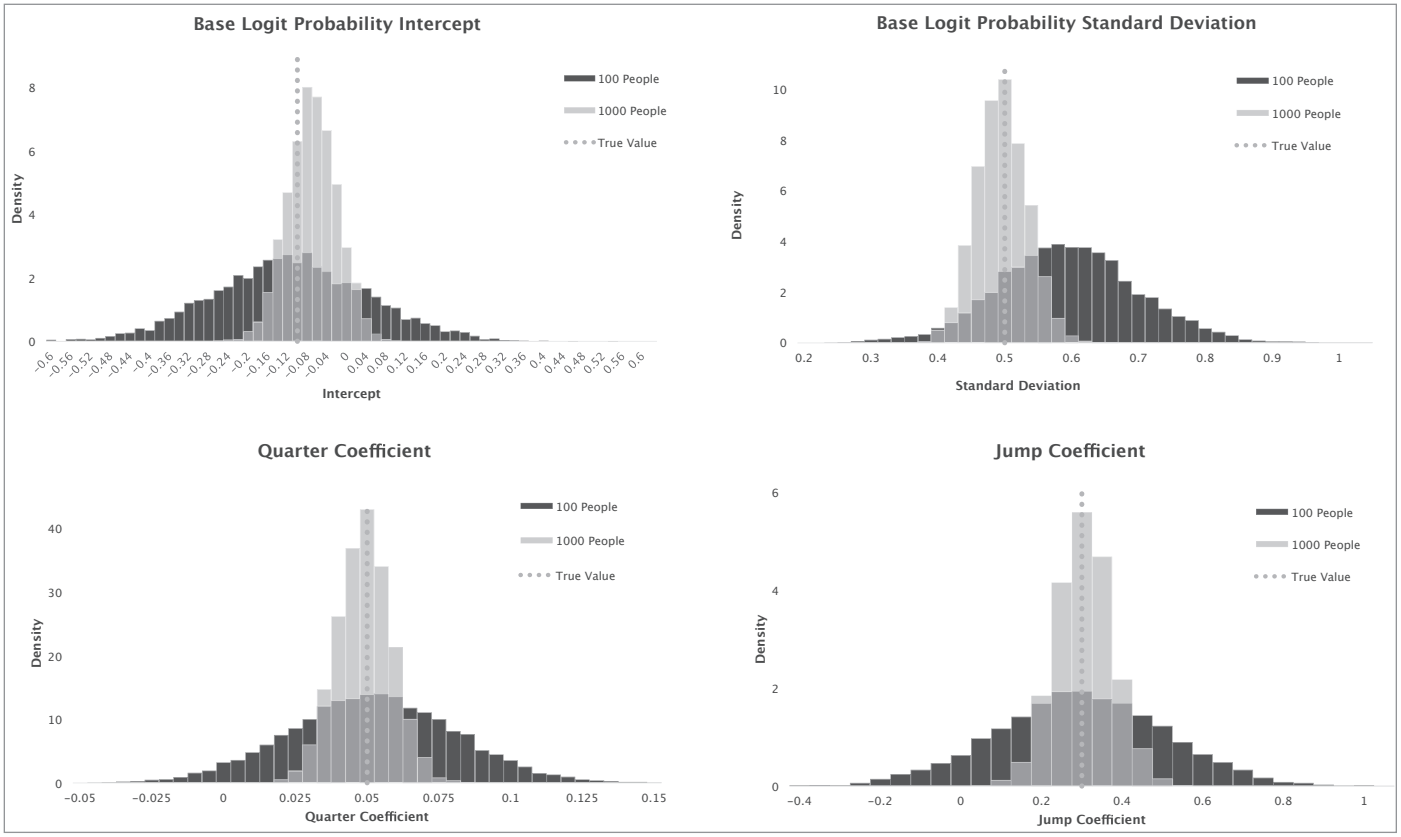
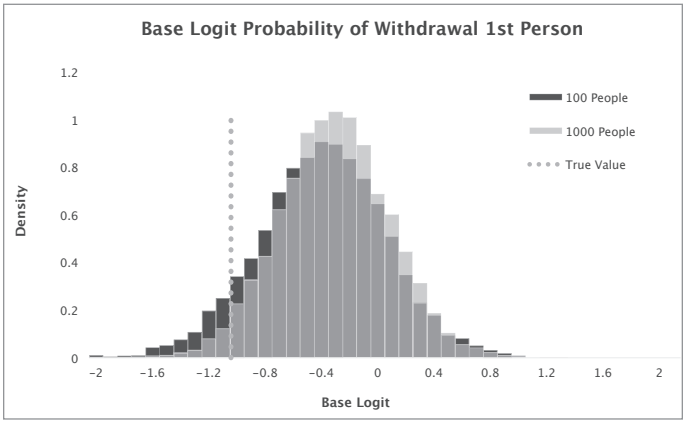
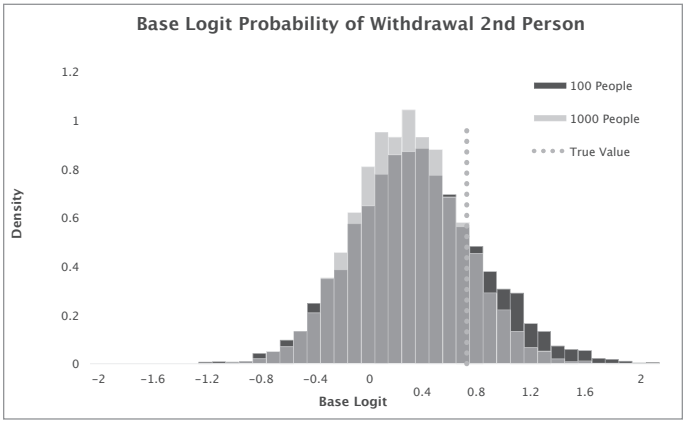


Figure 3
Inferred Distributions for Base Logit Probabilities for the First Two People in the Sample





Inferred PDFs of the four coefficients, together with their simulated true values, are shown in Figure 2. As one can see, having more data helps in the model coefficients inference—distributions become narrower.

Also, we can infer the base logit withdrawal probabilities for individual people—their withdrawal logit probabilities in the first quarter. The inferred base logit probability PDFs for the first two people in the samples are shown in Figure 3. These are the same two people from Figure 1. The two means that both sample PDFs in Figure 3 are very close, because they are the same two individuals in the two samples. These particular people have the same observations in 100- and 1,000-people samples, in terms of both total number of withdrawals and the withdrawal pattern from Figure 1. The true values in Figure 3 are a bit off from the means, because the first person had more withdrawals than expected and the second person had slightly fewer withdrawals than expected. But these true values are still within the posterior PDFs.

As one can see in Figure 3, we do infer higher values of the base withdrawal probability for the second person than for the first person, as we observed more withdrawal events for the second person. However, the distributions are fairly wide, because we have only 10 observations for these people. We can see that using 1,000-people sample makes inferred distributions a little narrower, because we have much better inferred model coefficients. However, even if we knew those coefficients exactly, the base logit distribution for individual people would still be fairly wide, because we have only 10 observations. Thus, we conclude that, for a good inference of the individual base probability, we need more longitudinal data—more quarters.

CONCLUSION

In this article we briefly described both MLE and Bayesian approaches in machine learning, looking at their advantages and disadvantages. We then proceeded with an example toy model that may be applicable for studying VA policyholder behavior. We used a simulation so that we fully understand the input data and the underlying true model.

For Bayesian inference, we used Stan probabilistic language. All inferred distributions made sense. As the amount of training data increases, the inferred distributions become narrower and closer to the true values.

This may serve as a good example for the reader in how to use Bayesian inference. ■

Denis Perevalov is a portfolio research analyst at Milliman in Chicago. He can be reached at *denis.perevalov@milliman.com*.

ENDNOTES

1 Pythonic Perambulations, "Frequentism and Bayesianism: A Practical Introduction" (March 11, 2014), *http://jakevdp.github.io/blog/2014/03/11/frequentism-and-bayesianism-a-practical-intro/*.

2 Frequentism and Bayesianism, ibid.

3 For more information, see *http://mc-stan.org/*.

4 See *https://github.com/Denisevi4/BayesianInference*.

REFERENCES

"Bayesian Inference in Machine Learning," *https://github.com/Denisevi4/BayesianInference*.

"Frequentism and Bayesianism: A Practical Introduction," *http://jakevdp.github.io/blog/2014/03/11/frequentism-and-bayesianism-a-practical-intro/*.

Stan, http://mc-stan.org/.

Maximal Information Coefficient: An Introduction to Information Theory

By Bryon Robidoux

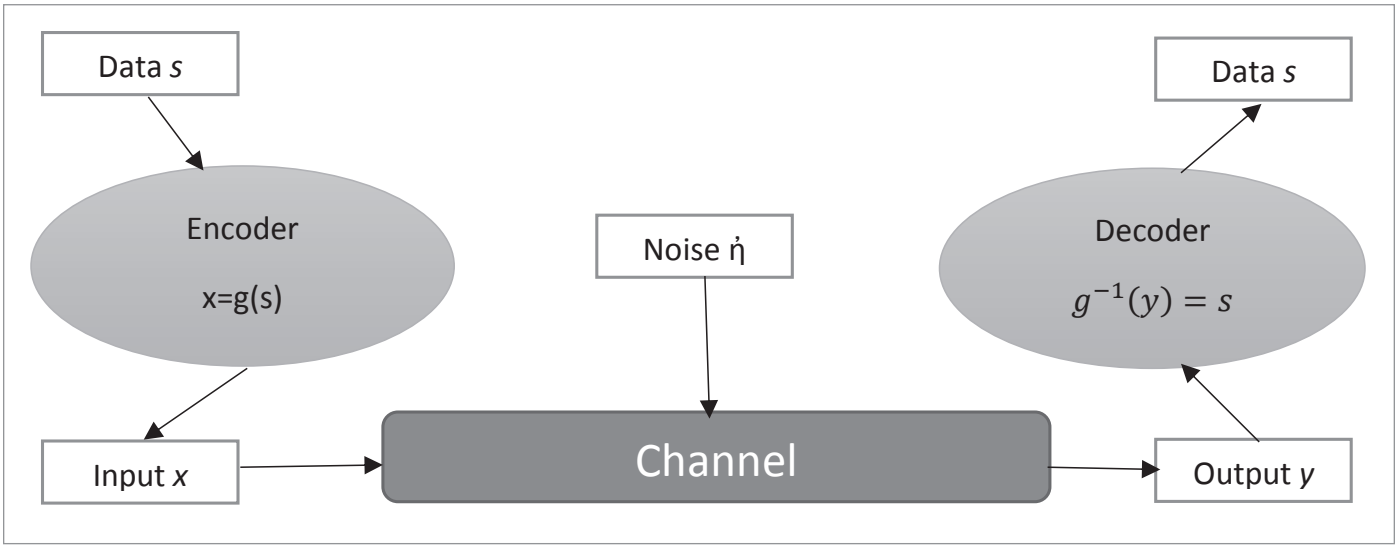
The maximal information coefficient (MIC) has been described as a 21st-century correlation that has its roots in information theory.[1] Information theory was developed by Claude Shannon back in 1948 when he published the paper "A Mathematical Theory of Communication" while working for Bell Labs. Scientists were trying to understand the limits of communication through a communication channel and how to send a signal and minimize the errors in the received message.[3] Even though this seems far removed from any problem in actuarial science, it turns out that it can be very useful for actuaries, such as:

1. Choosing between competing models for a stochastic phenomenon under investigation;
2. Adjusting mortality tables in a statistically valid manner to obtain exactly certain known or assumed individual characteristics, while simultaneously developing a table that is as close as possible to a given standard mortality table;
3. Smoothing observed insurance data to obtain smoothed estimates that are as close as possible to the observed data; and
4. Incorporating monotonicity constraints into a life table graduation.[4]

This article will describe the basic mechanics behind information theory, such as bits, entropy and mutual information, give some intuitive interpretation of its results and relate Pearson's correlation to MIC.

The basic unit of information theory is the bit, which stands for binary digit. This is unfortunate because a binary digit and a bit are different. A binary digit is the value of a binary variable, which can have only two values: zero and one. A bit is the amount of information required to choose between two equally probable alternatives. If there are m equally probable alternatives that can be arrived at by successively making n binary choices, then $n = \log_2 m$ bits of information are required. If the log is changed from base 2 to base e or base 10, then the units are nats or bans, respectively.[3] Information theory's original intent was to determine how to efficiently communicate information from point A to B with the least amount of error. Figure 1 shows the basic structure of communication.

Figure 1
Basic Structure of Communication



1. A source *s* generates a message, which is an ordered sequence of *k* symbols $s = (s_1,\ldots,s_k)$.
2. The source can be coded from an alphabet A_s which can have α letters, so $A_s = (s_1,\ldots,s_\alpha)$.
3. A message *s* is encoded as an input *x* by some function *g* into code words $x = (x_1,\ldots,x_n)$.
4. These code words can have their own alphabet with m letters, hence $A_x = (x_1,\ldots,x_m)$.
5. The input *x* is transmitted through the channel where noise ή is added to the output $Y = X + \acute{\eta}$.
6. The output y code words are decoded back into the original message.

Both the input *X* and output *Y* code words are defined as random variables, so there is a probability associated with each one of the encoded and decoded code words. The probability *p* of all the possible letters in an alphabet need to sum to unity. The output may not be the same as the input because the noise could have added errors into the transmission and changed the resulting alphabet character. The encoder is responsible for compressing and adding error-detecting redundancy. The decoder is responsible for decompressing the message and using the redundancy to remove errors from the message. The error rate in the transmission is the number of incorrect inputs associated with the output per the number of possible inputs.[3] Now that the original problem has been explained, it is time to formally define information and entropy.

Suppose that a biased coined is flipped and the probability of a heads is 95 percent. When the coin comes up heads, there is little information provided or surprise in this result. But if the result is tails, this is a lot more surprising and informative. The Shannon information is the amount of uncertainty or surprise in a random variable. It is defined as the $\log_2 1/p(z)$ bits, where *z* is any random variable, so the uncertainty of a variable should decrease as the probability of an event increases. The entropy $H(Z)$ is the expected value of the Shannon information $H(Z) = -\sum_i p(z_i)\log_2 p(z_i)$. A variable with $H(Z)$ bits entropy will have enough Shannon information to choose between $2^{H(z)}$ equally probable outcomes.[3] The calculation for the entropy is different for discrete versus continuous random variables. To see the problem, the entropy differential $H(Z^\Delta)$ needs to be defined: $H(Z^\Delta) = \sum_i p(z_i)\Delta z \log_2 \frac{1}{p(z_i)\Delta z}$. It is obvious that as $\Delta z \to 0$ then $H(Z^\Delta) \to \infty$. This can be interpreted as saying that as the precision of a variable increases, so dos the bits of information provided by the variable.[3] This means that integrals cannot be used to calculate the continuous entropy. To do the calculation, the random variables must be discretized by dividing the ranges into variable bins and counting how many values fall in the histogram grid.[1] Entropy has some very nice properties regardless if discrete or continuous:

- Continuity—the amount of information associated with an event increases or decreases continuously;
- Symmetry—the amount of information associated with a sequence of events does not depend on the order in which they occurred;
- Maximal Value—the amount of information associated with a set of events cannot be increased if the events are equally likely;
- Additive—the information associated with a set of events is obtained by adding the events together;
- Positive—it will always be greater than equal zero.[3]

The conditional entropy $H(Y|X)$ is the average amount of uncertainity in *Y* given that *X* has occurred, or, to phrase it another way, it is the amount of uncertainty in *Y* that cannot be contributed to *X*.[3] If the focus is put back on signal processing then the output *Y* is nothing more than the input *X* + random noise. The $H(Y|X) = H(X + \acute{\eta} \mid X) = H(\acute{\eta})$ so the average uncertainty in the output given the input is equivalent to the average uncertainty in the noise.[3]

The **relative entropy** between two distributions can be calculated by the Kullback-Liebler (KL) divergence. The relative entropy is a measure of the dissimilarity between probability distributions *p* and *q*: $KL(p||q) = \sum_k^K p_k \log_2 \frac{p_k}{q_k} = \sum_k^K p_k \log_2 p_k - \sum_k^K p_k \log_2 q_k$ where $\bullet_k^K p_k \log_2 q_k$ is called the **cross entropy**. The cross entropy is the average number of bits needed to encode data coming from a source with distribution *p* when model *q* is used. The KL divergence is the average number of extra bits needed to encode the data, due to the fact that the distribution *q* was used to encode the data versus *p*: $KL(p||q) \geq 0$ unless $p = q$.[1] Note that in general the relative entropy is not symmetric under interchange of the distributions *p* and *q*: in general $KL(p||q) \uparrow KL(q||p)$, so *KL*, although it is sometimes called the "KL distance," is not strictly a distance. The relative entropy is important in pattern recognition and neural networks, as well as in information theory.[2]

If there was a goal to state how one variable depends on another, one measurement that would suffice is to calculate the Pearson's correlation ρ that we are all so familiar with: $cov_{xy} = \sum_i^N \frac{(x_i - \bar{x})(y_i - \bar{y})}{N-1}$ *and* $\rho = \frac{cov_{xy}}{\sigma_x \sigma_y}$. Pearson's correlation measures only the linear relationship between random

variables x and y within a range [–1, 1] where 0,–1, 1 implies no relationship, perfectly negative relationship and perfectly positive relationship, respectively. Even though 0 implies no relationship, it does not imply that the random variables are independent. This is a limiting measure of dependence because many relationships are nonlinear. Mutual information is a more general approach of calculating how random variables depend on each other. It has a range from [0, ∞). There are actually several different formulas with corresponding interpretations for mutual information:

1. $I(X,Y) = KL(p(x,y) \| p(x)p(y)) = \sum_x \sum_y p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$.
   This is the extra bits needed to encode the data given independent distributions were used versus the joint distribution of X and Y.
2. $I(X,Y) = H(X) + H(Y) - H(X,Y)$. This is the intersection between the average uncertainty of the input and output.
3. $I(X,Y) = H(X) - H(X \mid Y)$. This is the difference between the average uncertainty of the input and the average uncertainty of the input knowing the output.
4. $I(X,Y) = H(Y) - H(Y \mid X)$. This is the average uncertainty of the output less the average uncertainty of the output given the input.
5. $I(X,Y) = H(Y) - H(noise)$. This is the difference between the average uncertainty of the output and the noise.[3]

Given that the mutual information is derived from the entropy, it suffers from the same problem of being infinite for continuous variables. Unfortunately, the number of bins used, and the location of the bin boundaries, can have a significant effect on the results of MIC. The maximal information coefficient is an approach to try many different bin sizes and locations, and to compare the maximum mutual information received. It is defined as $MIC \triangleq max_{x,y:xy<B} m(x,y)$ such that $m(x,y) = \frac{max_{G \in G(x,y)} I(X(G), Y(G))}{log_2 \min(x,y)}$ where B is some sample-size dependent bound on the number of bins that can be used to reliably estimate the distribution and $G(x,y)$ is the set of two-dimensional grids of size $x \times y$ and $X(G)$, $Y(G)$ represents a discretization of the variables onto this grid. The MIC lies in a range [0,1], where 0 represents no relationship between variables and 1 represents a noise-free relationship of any form, not just linear. MIC will not give any indication of the type of the relationship, though. It is possible with the MIC to find interesting relationships between variables in a way that simpler measures, such as the correlation coefficient, cannot.[1] With MIC the goal is equitability: similar scores will be seen in relationships with similar noise levels regardless of the type of relationship. Because of this, it may be particularly useful with high dimensional settings to find a smaller set of the strongest correlations. Where distance correlation might be better at detecting the presence of (possibly weak) dependencies, the MIC is more geared toward the assessment of strength and detecting patterns that we would pick up via visual inspection.[5]

In conclusion, this article has taken you from the elementary beginnings of information theory. The concepts of bits and nats were explained, which led to the definition of entropy and its many flavors as well as the definition of the KL divergence and its interpretation. This led to the description mutual information for the discrete and continuous cases. Last, the familiar Pearson's correlation coefficient was compared to MIC. MIC is important to pattern recognition because it is a general approach to measure the dependency between two random variables, whereas Pearson's correlation measures only linearity between two random variables. ■

Bryon Robidoux, FSA, is director and actuary at AIG in Chesterfield, MO. He can be reached at *Bryon.Robidoux@aig.com.*

ENDNOTES

1 Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.

2 MacKay, David J. C. 2003. *Information Theory, Inference, and Learning Algorithms,* 2nd ed. Cambridge: Cambridge University Press.

3 Stone, James V. 2015. *Information Theory: A Tutorial Approach*. Sheffield, UK: Sebtel Press.

4 Brocket, Patrick L. 1991. "Information theoretic approach to actuarial science: A unification and extension of relevant theory and applications," *Transactions of the Society of Actuaries,* vol. 43.

5 Clark, Michael. 2013. "A comparison of correlation measures," Center for Social Research, University of Notre Dame, *https://m-clark.github.io/docs/Correlation-Comparison.pdf.*

A Global Qualification
Certified Actuarial Analyst (CAA)

Benefit from enhanced technical and analytical skills

With the mathematical skills and ability to communicate financial concepts, manipulate and analyze data sets, CAAs are essential members of your actuarial team.

The CAA is offered by CAA Global, a joint venture of the Institute and Faculty of Actuaries and the Society of Actuaries.

Learn more at CAA-Global.org and continue building to the future.

Variable Selection in Predictive Modeling: Does it Really Matter?

By Kailan Shang

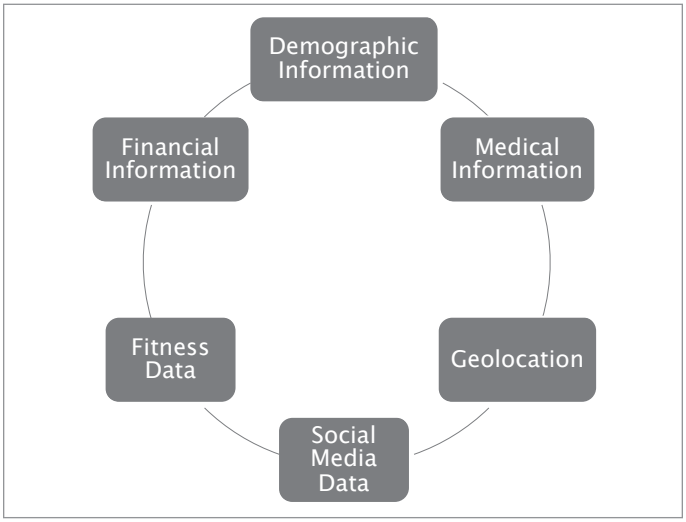
Many actuarial works have been expanded in the era of big data. Risk analyses are moving from the aggregate level to the individual level enabled by better data availability. For example, mortality risk can be assessed not only by traditional data such as age, gender, smoker/nonsmoker, occupation, face amount and basic medical information, but also new data, including location, detailed medical information, financial status, fitness data and even social media data. These new data sources can help us learn more about individuals or events that affect the mortality trend. In addition, some new data are categorical and cannot be used directly by predictive models like numerical data. For example, cancer patients have different tumor sites and medical treatments. An insurance client may participate in different types of sports. One categorical variable could become dozens of numerical variables, with each indicating the presence of a specific variable. The number of explanatory variables could easily exceed a few hundred.

DO WE NEED VARIABLE SELECTION?

With so many variables, is it necessary to select a subset of variables with the best performance of prediction? For traditional predictive models used by actuaries, the answer is obviously positive. The robustness of linear regression models and generalized linear models (GLMs) can be low with the presence of collinearity caused by too many variables. The prediction results will be very sensitive to the input data. However, some machine learning models such as random forests and artificial neural networks (ANNs) were designed to handle large data input without prior assumption of the data relationship. Dimension reduction techniques such as principal component analysis (PCA) and autoencoder could also systematically reduce the number of explanatory variables. The needs for variable selection are less obvious for these models.

However, the benefits of variable selection go beyond model training and model selection. By selecting the best predictors, people can understand the most important relationships implied from the data. It is easier for people to assess these relationships at a small scale rather than being overwhelmed with hundreds of variables at the same time. Reducing the number of explanatory variables also decreases the chance of overfitting. Overfitting happens when too many variables are unintentionally used to explain the random noises instead of the relationships. The variance of prediction is large even though the accuracy of prediction may be high for the training data. Figure 2 illustrates an example of overfitting. A linear model with one explanatory variable X1 could explain the main relationship even though its accuracy is lower than a perfect matching nonlinear model with much more explanatory variables.

Figure 1
Data Sources for Individual Mortality Prediction



Overfitting can be overcome by analyzing the contribution of each variable to the prediction and removing variables with trivial contributions. Variable selection may not improve the model accuracy measured by the training data, but it can certainly improve the robustness of found relationships. Maintaining only the important variables in the predictive models also helps explain the model. The application of the model to new data will be more efficient. Less data collection, storage and calculation can be achieved by variable selection.

On the other hand, variable selection is challenging for big data. Will predictive models be able to identify important variables automatically? The answer is both yes and no. Predictive models are instrumental for identifying useful variables, but they are not always working in a desired way.

USING PREDICTIVE MODELS

A few approaches can be used to select important variables by running multiple models. The forward approach starts from an empty model and adds one variable at a time. At each step, the variable with the biggest accuracy improvement is chosen. The

Figure 2
Overfitting Illustration

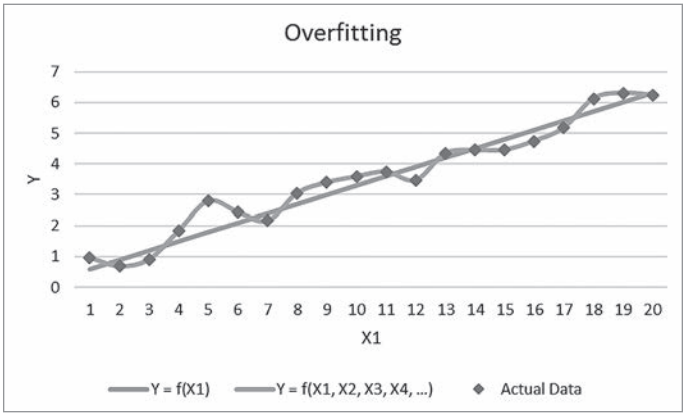
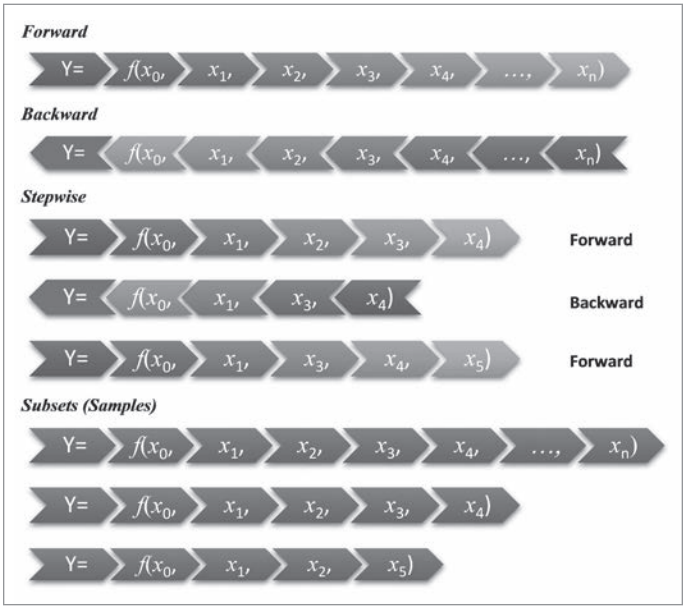


Figure 3
Variable Selection Methods



forward process ends when the model accuracy stops improving or the improvement is trivial. The backward approach starts from a full model with all variables and removes one variable at a time. At each step, the variable with the biggest negative impact or the smallest positive impact is removed, until the model accuracy stops improving or reaches the desired level. However, the problem with both the forward and backward approach is that the sequence of the explanatory variables matters. Adding a new variable to the model could change the importance of existing variables. The stepwise approach addresses this issue by combining the forward approach and the backward approach. At each step, an additional variable is added, and then the new model works backward to remove any existing variables that have a negative or trivial impact on model accuracy. Another more comprehensive yet costly approach is to iterate through all possible combination of explanatory variables and choose the subset with the smallest set of variables given that the model accuracy meets the target.

When applying these approaches, many measures can be used to represent model accuracy. The measures are used in two places: the target above which variable selection process will stop and the minimum positive improvement deciding whether a variable should be added or dropped. Table 1 lists a few measures for regression and/or classification models.

However, these four approaches are expensive given the number of models that need to be run. It could be very challenging for big data with many variables. Table 2 lists the maximum number of models that need to be trained to finish the variable selection process for each approach assuming n explanatory variables. The actual number of models could be smaller than the maximum number because the process could stop once the target accuracy is achieved.

To reduce the burden of additional model training, variable selection can be done based on the result of the complete model with a couple of adjustments. After the model is trained, the importance of each variable can be measured to determine its contribution to the prediction. Variables are then selected based on their importance. Several adjustments to the modeling process can be made to address the issue of overfitting in one model training:

1. Collinearity/multicollinearity checking. Variables with high correlation, either positive or negative, can be reduced. If the absolute value of correlation coefficient exceeds a threshold such as 95 percent, one variable of the pair can be removed. For multicollinearity where one explanatory variable can be explained very well by other explanatory variables, the explanatory variable can be removed as well because its information can be provided by the remaining variables. Multicollinearity can be assessed using the variance inflation index (VIF). For an explanatory variable x_i, a linear regression can be run against other explanatory variables:

$x_i = \alpha + \beta_1 x_1 + \quad + \beta_{i-1} x_{i-1} + \beta_{i+1} x_{i+1} + \quad + \beta_n x_n$. Its VIF is calculated as

$$\frac{\sum_{j=1}^{m} \left(x_i^j - \overline{x_i} \right)^2}{\sum_{j=1}^{m} \left(\hat{x}_i^j - x_i^j \right)^2}$$

Table 1
Variable Selection Measures

Akaike Information Criterion (AIC)	$2p - 2\log(likelihood)$	All
Bayesian Information Criterion (BIC)	$plog(m) - 2\log(likelihood)$	All
Adjusted R^2	$1 - \frac{\sum_{i=1}^{m}(\hat{Y}_i - Y_i)^2 / (n-p-1)}{\sum_{i=1}^{m}(Y_i - \bar{Y})^2 / (n-1)}$	Regression
Mean Square Error (MSE)	$\frac{1}{m}\sum_{i=1}^{m}(\hat{Y}_i - Y_i)^2$	Regression
Mean Absolute Error (MAE)	$\frac{1}{m}\sum_{i=1}^{m}\lvert Y_i - Y_i \rvert$	Regression
Precision	$\frac{True\ Positive}{True\ Positive + False\ Positive}$	Classification
Recall	$\frac{True\ Positive}{True\ Positive + False\ Negative}$	Classification
F-Measure	$\frac{2 \bullet Precision \bullet Recall}{Precision + Recall}$	Classification

Notes:
p: number of variables
m: number of data records
Y_i: actual value of explained variable for the ith data record
$\hat{Y}_i$: predicted value of explained variable for the ith data record
$\bar{Y}_i$: average value of the explained variable

Table 2
Models under Four Variable Selection Approaches

Maximum No. of Models	$\frac{n(n+1)}{2}$	$\frac{n(n+1)}{2}$	$\mathbb{O}(n^3)$	$2^n - 1$

where

m: number of data records

x_i^j : the value of x_i for the *j*th data record

$\overline{x_i}$: the average value of x_i for the *m* data records

$\hat{x}_i^j$: the predicted value of x_i based on other $(n-1)$ explanatory variables for the *j*th data record.

Kutner et al. (2004: 408–409) suggest that a VIF greater than 10 or the mean value of VIF for all explanatory variables greater than 1 indicates the existence of multicollinearity.

2. Data normalization. To facilitate variable importance measurement, explanatory data can be normalized into the same value range. By doing this, variable importance can be determined by the magnitude of the model parameter for that variable. For example, in an linear equation such as $Y = 0.5 + x_1 + 4x_2$. If both x_1 and x_2 are within the same value range, we may simply conclude that x_2 is four times more important than x_1 in the prediction. For nonlinear relationship, it is more complicated but normalization is still useful for a consistent comparison. Normalization can be done in different forms such as feature scaling and standard score:

   Feature scaling: $\frac{X - X_{min}}{X_{max} - X_{min}}$

   Standard score: $\frac{X -}{\sigma}$

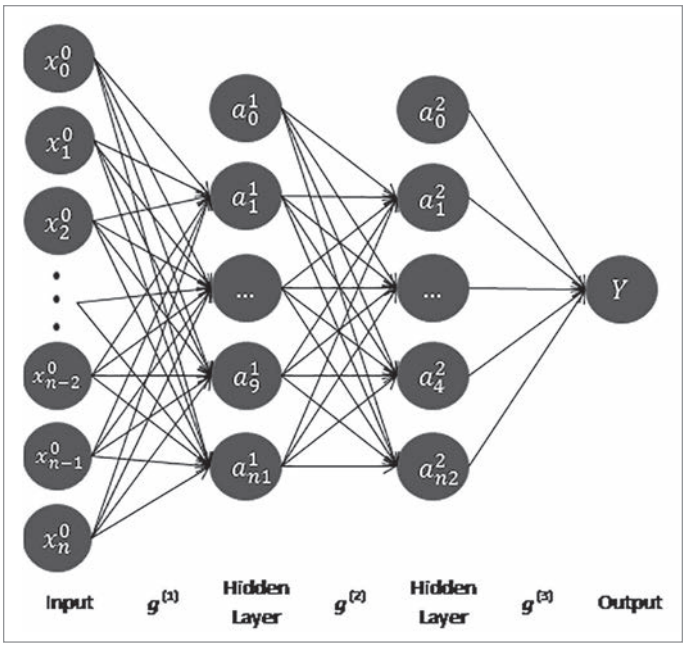
3. Regularization is often used in models that can handle many variables to address the issue of overfitting. By introducing the penalty for model complexity, it does not explicitly select variables in the model but limits the value of model parameters. For example, ridge regression intends to minimize the sum of squared errors and squared parameters. Parameter λ controls the weight of the penalty:

$$\min_{\beta} \sum_{j=1}^{m} \left(Y_j - \sum_{i=1}^{n} x_i^j \beta_i \right)^2 + \lambda \sum_{i=1}^{n} \beta_i^2$$

Normal regularization includes L1 regularization, which uses the sum of the absolute value of parameters, and L2 regularization, which uses the sum of the squared value of parameters, as in the ridge regression. Models such as random forest do not have model parameters for each variable. Other approaches are used for regularization such as controlling the maximum depth of the trees to avoid overfitting.

After all these adjustments, variable importance can be measured and used for variable selection. For model with normalized data, the absolute value of coefficients can be used for models like linear regression and GLMs to determine the relative importance of variables. For more complicated models, the calculation of relative importance is more complicated. For example, for an ANN model with two hidden layers, the impact of the explanatory variables is determined through three sets of parameters: $g^{(1)}$, $g^{(2)}$ and $g^{(3)}$, as illustrated in Figure 4.

Figure 4
ANN Model Structure



A possible measure is to consider the impact of the explanatory variable through three layers, including the two hidden layers and the output layer:

$$Imp(x_i) = \sum_{j=1}^{n1} \sum_{k=1}^{n2} \frac{|\theta_{ij}^0|}{\sum_{r=1}^{n} |\theta_{rj}^0|} \cdot \frac{|\theta_{jk}^1|}{\sum_{s=1}^{n1} |\theta_{sk}^1|} \cdot \frac{|\theta_{kY}^2|}{\sum_{t=1}^{n2} |\theta_{tY}^2|} \quad (19)$$

where

x_i : The *i*th input variable

$n1$: The number of neurons in the first hidden layer

$n2$: The number of neurons in the second hidden layer

n : The number of explanatory variables

θ_{ij}^0 : The parameter that determines the weight of the *i*th input variable applied to the *j*th neuron in the first hidden layer

θ_{jk}^1 : The parameter that determines the weight of the *j*th neuron in the first hidden layer applied to the *k*th neuron in the second hidden layer

θ_{kY}^2 : The parameter that determines the weight of the kth neuron in the second hidden layer applied to the output variable Y

This measure also has its disadvantages because it cannot tell whether the relationship is positive or negative. It also does not consider the specific function used to link the layers.

For tree-type models like random forests, the measurement of variable importance is different and even more complicated. A possible measure is the Gini importance measured by the improvement of the Gini impurity index. The Gini index is defined as

$$G(T) = \sum_{i=1}^{n} p_i (1 - p_i)$$

where

p_i is the probability that the data belongs to category i

n is the number of categories in the data

T is the data set based on which Gini index is calculated

For each split based on the variable, the Gini importance is measured as the reduction in the Gini index:

$$Imp(x_i) = n(T)G(T) - n(T_L)G(T_L) - n(T_R)G(T_R)$$

where

x_i is the variable for the split

T_L is the data subgroup of the split's left branch

T_R is the data subgroup of the split's right branch

n is the number of data points in the data set

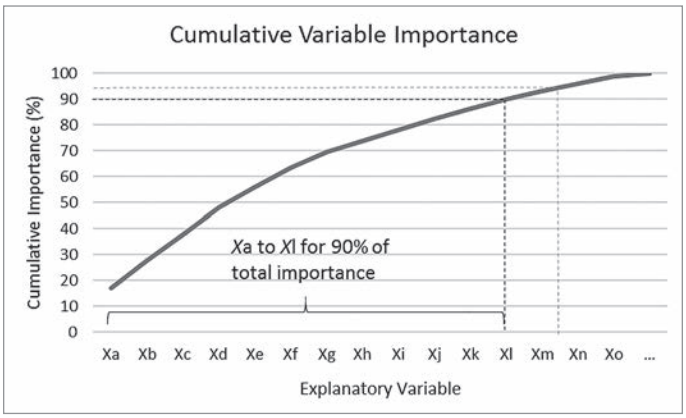
p is the portion of the data subgroup in the data set before splitting

If the variable is used in multiple splits, the Gini importance is aggregated for the variable. For random forests with multiple trees, the mean Gini importance across all trees can be used to measure variable importance.

Permutation importance can also be used to measure variable importance in tree models. The prediction is revised by permutating the value of the variable, and the loss of prediction accuracy is used as the importance measure for that variable. When variables are highly correlated, conditional permutation can be used to maintain the correlation. However, this is less of a concern after collinearity/multicollinearity checking.

After the variable importance is calculated, the top variables can be selected for future prediction. The threshold can be set based on a specified portion of total importance that selected variables explain in aggregate. Figure 5 illustrates the variable selection based on the cumulative variable importance. Variable importance is scaled so that the total importance is 100 percent.

Figure 5
Variable Selection Based on Variable Importance



THE ROLE OF EXPERT OPINION

Although using predictive models to automatically search for important variables is a convenient and consistent approach, human judgments are needed at various stages of the process. At the initial stage, explanatory variables need to be screened one by one to assess their relevance to the explained variable. Both a blacklist and a whitelist of the variables can be created. If strong evidence exists for the irrelevance of an explanatory variable, the variable can be added to the blacklist and removed from the entire process. On the contrary, for variables that are believed to have a strong relationship with the explained variable, they can

be added to the whitelist and kept in the model. In the collinearity analysis, when a pair of variables are found highly correlated, human judgment is also needed to decide which one is more likely the root cause and should be retained in the analysis.

After the variable selection finishes, the reasonableness of the relationships derived from the data needs to be assessed. Sometimes even if the model accuracy is satisfactory, the relationship could be inconsistent with past experience, scientific findings and common sense. Additional work needs to be done to before accepting or rejecting the relationships. More data collection, model adjustments and different variable selections could be triggered by human judgment.

CONCLUSION

Although many models can address the overfitting issues caused by too many variables by regularization, variable selection is still meaningful. The model and data are more parsimonious, and it is easy for people to assess, understand and explain the relationships derived from the data. Variable selection can be done through either multiple models or measures based on the complete model with adjustments. Measures might be complicated and different depending on the model, but they are computationally cheaper than multiple model runs. Human judgment is also important in the process of variable selection to incorporate expert opinions based on existing knowledge and experience.

Kailan Shang, FSA, CFA, PRM, SCJP, is managing director of Swin Solutions Inc. in Kitchener, Ontario. He can be reached at *kailan.shang@swinsolutions.com.*

REFERENCE

Kutner, Michael H., Christopher J. Nachtsheim, John Neter and William Li. 2004. *Applied Linear Regression Models*. New York: McGraw-Hill.

The First SOA Annual Predictive Analytics Symposium—A Recommended Investment! (Whether or Not Your Employer Pays for It)

By Dave Snell

Think back to when you decided to become an actuary. The required education was probably a huge time and dollar commitment, but most of us would agree that it was an excellent investment.

Actuaries rank near the top of most lists for the best job in the U.S. We enjoy a profession that offers high wages, pleasant and nontoxic working conditions, and a lot of satisfaction that what we do helps millions of families enjoy a more secure and enjoyable future.

However, we can't afford to be too complacent! A new wave of professionals, the data scientists, are pushing actuaries from those coveted top spots on the Best Jobs lists. In fact, the *Harvard Business Review* called Data Scientist "The Sexiest Job of the 21st Century" (October 2012). The good news, though, is that you can be both a data scientist and an actuary—a match that can ensure (or perhaps, insure?) your continued market value in a world becoming increasingly dependent upon predictive analytics.

We are all seeing flyers, emails and other advertisements for commercial conferences on predictive analytics (PA), big data, predictive models, data analytics and similar titles that promise wonderful returns if you attend them. Unfortunately, my colleagues and I come back complaining that we had to sit through dozens of sessions to find even one potential application to insurance. What if there were an entire conference, with dozens of sessions, all focused on ways to help you, as an actuary, capitalize on this PA explosion?

Now, there is—and it is sponsored by the Society of Actuaries! Furthermore, it was organized in collaboration with the Predictive Analytics and Futurism section (PAF) council and friends of the council.

This September 14–15, in Chicago, the SOA will host the first annual Predictive Analytics Symposium. It will have multiple tracks for PA so that you can choose to follow the management route, the beginning or intermediate practitioner route, or the advanced PA techie route. A manager might feel "I don't want to have to become a techie again, but I want to understand how PA can help my company in my specific interest area" (e.g., life, health, general insurance, ERM, etc.). A midlevel or newer actuary might want to dive in and become literate (or more knowledgeable) on the most cost- and time-effective ways to get productive with a classification and regression tree (CART) or a random forest. A person already using PA might want to learn the cutting-edge techniques (Deep Learning with Tensor Flow, advanced distribution choices, etc.). Alternatively, a person strong in one aspect of PA but wanting to delve into both the breadth and depth of PA can choose to mix and match throughout the conference.

Hopefully, your employer will see the tremendous value here and fund your trip to Chicago (easy access and not quite as pricy as some other areas of the country). But what if you have to pay for it yourself? I realize I am writing to a group of six-figure earners who sometimes balk at the $25 per year section membership if not paid by their employer, but let's get real here. This is a great investment—no matter who makes it.

You will come away with ideas and with immediate applications from peers across the globe who are focused on the same financial risks that you are. This is not a Predictive Analytics World (PAW) conference where you learn how to manufacture a part or design an autonomous car. It is strictly for insurance and related financial risk applications.

The keynote speakers will be actuaries or chief data analytics officers from insurance companies you know and respect. The sessions will be conducted by actuaries who have learned to be successful in the PA space or data scientists who work with actuaries. Several of the speakers will be PAF section members who have utilized PA for specific insurance applications.

We are planning about three dozen sessions, arranged in three (or four) concurrent tracks. These will cover introductory topics such as how PA can help you in life, health, general insurance and other specialty areas (you attend the sessions that interest you), and progress all the way to the cutting edge of Deep Learning Neural Networks (how to use TensorFlow to write your own applications) and the wonders of machine learning. They should appeal to the manager who wants to know how to build a data analytics team, to the actuary who wants to move into this exciting new area (at various levels: running models, building models, managing others who run or build models) and to experienced practitioners who want to learn the latest and greatest extensions to further their expertise. We are bringing this all together under one roof, so to speak, so that you can find and network with actuaries who share interests in very special application areas.

This issue is timed to reach you a month earlier than before (June instead of July), and it should coincide with the registration information and long list of session titles and descriptions. Please check for the announcement from the SOA. Our section is providing a large group of presenters, and they are being supplemented with noted experts from North America and Europe (so far . . . we are still recruiting more presenters) to bring actuaries together for a PA experience specifically for actuaries. My first draft of this article went for six pages of session titles and descriptions, but it would have been out of date by the time it reached you. Please check out the conference details when you receive them.

Oh, yes, and one more thing: this is not meant to be a substitute for the SOA Annual Meeting, the Health Meeting, Life and Annuity Symposium, Valuation Actuaries Meeting or any of the other fine conferences you may attend. We want you to supplement whatever education you have been experiencing with something new and exciting and immediately useful that will put the actuary back on top of the best jobs lists. Take note, *Harvard Technology Review:* The actuaries are taking over the PA space for financial services! ■

Dave Snell, ASA, MAAA is technology evangelist at SnellActuarialConsulting in Chesterfield, Mo. He can be reached at *dsnell@ActuariesAndTechnology.com*

SOCIETY OF ACTUARIES

475 N. Martingale Road, Suite 600
Schaumburg, Illinois 60173
p: 847.706.3500 f: 847.706.3599
w: www.soa.org

NONPROFIT
ORGANIZATION
U.S. POSTAGE
PAID
SAINT JOSEPH, MI
PERMIT NO. 263