Article from

Predictive Analytics and Futurism

June 2017
Issue 15

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.

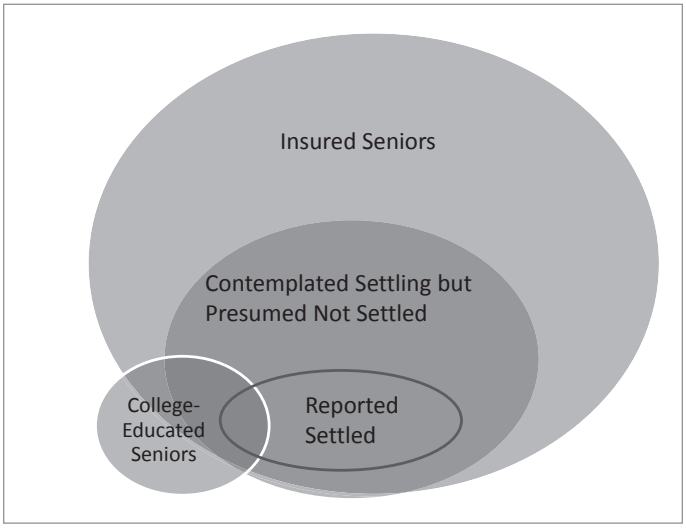
THE OBJECTIVE

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

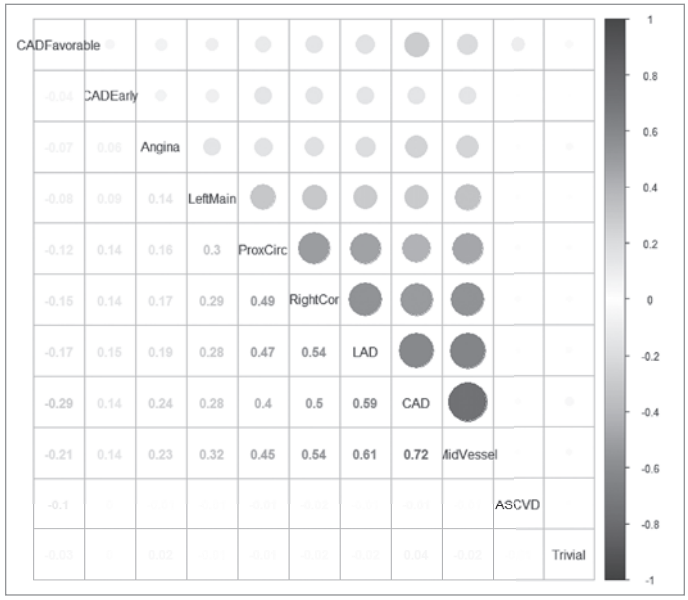
Exhibit 1 illustrates the output of the current Cox Proportional Hazards model for the CAD and Coronary Anatomy sections. Besides the name of the condition, we included a count, the number of underwritings where the subject was found with the condition, the log of the hazard, the hazard ratio (the mortality 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.*