July 2023
ReCent
Actuarial News
somewhat different

Beyond actuarial analysis: Towards a systematic way of approaching risk

Background

In a competitive life insurance market, accurate risk pricing is vital to attract new policyholders without being selected against. Traditional actuarial experience analysis can give valuable insights into the biometric risk that an insurer is exposed to, and the trends in this experience over time. However, there are some limitations with traditional analysis, and more cutting-edge data science techniques can offer more insight to help understanding the risk.

Take the example of an experience analysis that shows a trend of improving biometric experience by underwriting year. One might conclude that more recent business is performing better than historic business, allowing new business pricing to be set at a level below the aggregate historic experience.

Perhaps underwriting standards have been improving over time, or marketing and distribution have driven lower-risk policyholders to favour your organisation. However, a similar trend might be seen if the experience is actually deteriorating by policy duration, because the earlier underwriting years are the policies that contribute to the later duration experience in your analysis.

Using traditional methods, it can be very difficult to tell these effects apart, and even more challenging to disaggregate the impact. However, more sophisticated models allow you to control for multiple risk factors and gain more insight into the true drivers of the experience. In the example above, it is possible to separate the impact of duration and underwriting year, and see the true underlying trends.

This does not negate the need to understand the business written and the drivers of the underlying experience. An understanding of the reason we observe an improving trend by underwriting year for example, is vital to ensure the right judgements are made in pricing new business, but new techniques are extremely useful in highlighting the key drivers of historic experience.

Pitfalls of traditional analysis

Traditional experience analysis is subject to a number of pitfalls, stemming from not considering data from a holistic perspective and information loss due to data transformation. These limitations fall into four categories discussed below.

Limitation 1: Not modelling continuous variables appropriately

Due to challenges in visualising relationships between continuous predictor variables (such as age) and outcomes, and to increase statistical power for each point estimate, continuous variables are frequently banded into groups. As discussed previously[1], this results in loss of information and can obscure patterns (such as the critical illness spike at age 50 amongst females due to breast cancer screening).

Loss of information can be twofold: first information is lost by combining different values into a single group and secondly, if the banded variable is treated as a category, the continuous sequence is lost to any subsequent model.

Limitation 2: Risk of misinterpretation due to confounding factors

Confounding occurs when the relationship between a variable and outcome is distorted due to the presence of a second variable that is associated with both the first

[1] https://www.hannover-re.com/1650132/recent-uk-insights-an-insight-into-accelerated-critical-illness-experience-2021.pdf

variable and the outcome. Confounding can result in misinterpretation of study results:

Misattribution of effects

If a confounding variable is not properly accounted for, confounding can create a spurious association or mask a real association. For example, failing to adjust for smoking (the confounder) may show a spurious association between coffee drinking and lung cancer because smokers are more likely to drink coffee than non-smokers.

Distorted association

Confounding can lead to overestimation or underestimation of the true association between the predictor variable and outcome. Consider the hypothetical dataset summarised in table 1.

Pricing males at 120%, females at 80%, joint lives at 80% and single lives at 120% (adjustment 1) will not price this dataset accurately even in the absence of an interaction (more on interactions in the next section) between gender and joint life status. The reason is that gender and joint life status are correlated risk factors in this dataset (table 2): most males are single lives and most females are joint lives. In adjustment 1, the effect sizes of gender and joint life status are overestimated due to confounding. In this specific example, pricing adjustments are relatively easy to make by directly calculating actual/expected ratios for each category individually (adjustment 2).

Predictor variables are rarely non-correlated. Manual adjustment for multiple correlated variables gets progressively more complicated, especially for numeric variables.

Table 1: Hypothetical dataset – by gender and joint life status univariate

Feature	Category	Actual	Expected	Actual / Expected ratio Standard table
Gender	Male	600	500	120%
	Female	400	500	80%
Joint life status	Joint	600	500	120%
	Single	400	500	80%

Table 2: Hypothetical dataset – by gender and joint life status multivariate

Gender	Joint life status	Actual	Expected	Adjustment 1	Adjustment 2
Male	Single	478	375	375 * 120% * 120% = 540	375 * 127.6% = 478.5
Male	Joint	122	125	125 * 120% * 80% = 120	125 * 97.3% = 121.6
Female	Single	122	125	125 * 80% * 120% = 120	125 * 97.3% = 121.6
Female	Joint	278	375	375 * 80% * 80% = 240	375 * 74.2% = 278.3
Full dataset		1,000	1,000	1,020	1,000

Limited generalisability

Confounding can impact the generalisability of study findings. If the confounding variable is unevenly distributed across study groups, the observed association may not be generalizable to other populations or contexts.

For example, sales channel pricing derived by a study confounded by socioeconomic factors will only be valid for populations with the same socioeconomic breakdown by sales channel.

Limitation 3: Little thought given to variable interactions

Variable interactions occur when one predictor variable modifies the effect of another variable on the outcome. For example, the select shape could hypothetically differ between different sales channels.

In these cases, pricing the select shape and sales channel effect accurately on the aggregate level will not take into account the interaction, which requires a different select shape for each sales channel. Variable interactions can only be detected by explicitly considering each pair of variables which can be very time consuming (limitation 4).

Some models such as tree-based models and deep learning models do inherently take variable interactions into account. The trade-off to this is the non-transparency of such models.

Limitation 4: Lack of systematic method for feature selection

In addition to being vulnerable to confounding, case-by-case analysis of variables can be time consuming. This is even more so if pairs (or even triplets) of variables are considered on a case-by-case basis to explore for interactions. For example, with 10 variables, there are 45 potential two-way interactions to consider.

Workflow for addressing limitations in traditional analysis

A possible solution to the issues described above is to incorporate a correlation matrix as well as an assessment of variable importance in traditional experience analysis. This is best incorporated into data exploration. Neither of these techniques requires the banding of continuous variables, preventing loss of information (limitation 1).

Correlation matrix

A correlation matrix is a table that systematically displays the correlation coefficients between variables in a dataset. It provides a comprehensive view of the strength and direction of pairwise relationships between variables.

The correlation matrix can:

1. Identify potential confounding variables, which can prevent misattribution of effects, distorted associations and limited generalisability (limitation 2).
2. Provide a deeper understanding of the dataset. For example, a change in business mix over time. This may even result in exclusion of certain data from analysis (for example a sales channel that is no longer providing new business) in order to improve generalisability.
3. Identify redundant variables and prevent multicollinearity. Strongly correlated predictor variables may be capturing the same characteristics and some could safely be removed from analysis without information loss. Avoiding strongly correlated predictor variables can also prevent issues with multicollinearity.
4. Identify potential data errors. Data entry errors / data missing not at random could cause unexpected strong correlations between variables. For example, rated lives not being reported for certain sales channels would result in an unexpected and strong correlation between rated lives and sales channel.

Variable importance

A variable importance table is a summary that ranks predictor variables in a model based on their importance or contribution to the model's performance. Variable importance provides a measure of the relative importance of each variable in predicting the outcome. Variable importance tables can be generated by various models including tree-based models such as random forest and gradient boosting machines as well as LASSO regression.

Variable importance tables provide a systematic method for assessing the relative importance of all predictor variables in a single (holistic) model (limitation 4). Additionally, all potential variable interactions can be explicitly generated using the model design matrix and considered using the variable importance table (limitation 3). The model underlying the variable importance table will also disentangle confounding effects (limitation 2) and assign importance to the root cause of effects. Variable importance tables provide a time efficient way of determining which predictor variables drive an outcome.

Downstream workflow

The correlation matrix and variable importance will provide systematic insights into what data, variables and variable interactions to consider for a production model. Incorporating the correlation matrix and feature importance into the data exploration workflow will greatly reduce the risk of confounding, unidentified variable interactions or poorly generalizable models.

Limitations

Confounding factors may not be captured in the dataset. For example, in the limited generalisability example of sales channel being confounded by socioeconomic factors, socioeconomic information may not be captured in the dataset. In such cases, data analytics will be unable to detect and adjust for the confounding variable. Expert opinion is the only tool available for these cases and feed into any analysis.

Although variable importance tables provide a systematic overview of which variables drive the outcome, feature selection itself should take place within the production model framework (for example by stepwise selection or LASSO regression). In case of a large amount of variables, any variables where variable importance scores similar or lower than random noise can be excluded during data exploration.

Conclusion

Incorporating a correlation matrix and variable importance calculation into the data exploration workflow of experience analysis is a time efficient way to greatly reduce the risk of confounding, missing variable interactions or poorly generalizable models.

Hannover Re use insights derived from the workflow described to set our best estimate view of mortality/morbidity rates and we can support more granular risk assessment/pricing. Hannover Re can also deliver client-specific insights using similar methods.

Authors

Michiel Luteijn Ph.D.
Senior Data Scientist
Hannover Re UK Life Branch
Tel. +44 20 3206 1828
michiel.luteijn@hannover-re.com

Tim Smith
Head of Protection
Hannover Re UK Life Branch
Tel. +44 20 3206 1811
tim.smith@hannover-re.com