Analysis Guidelines

The dataset to be analyzed is available on the Society of Actuaries (SOA) website here: http://cdn-origin.soa.org/research/2009-15_Data_20180601.zip. This file is a text file in tab-delimited format. This is a preliminary dataset that has been through validation and relatively extensive consistency checking, but has not been exhaustively analyzed, especially using some of the latest data science and predictive analytics techniques.

Entrants must analyze the data and determine what issues, gaps, inconsistencies, problems, outliers, etc., if any, exist with the data set. We have also supplied a document here that provides additional information about the fields contained in the database. Numerous fields – such as policy type – were supplied by the contributing companies and the external user would have no way of validating these entries. Some of the entries can be derived from others (but may be off by a year because of rounding). Other entries – such as exposures and deaths – reflect the actual experience submitted.

Entrants in the competition are expected to analyze the data and find issues, gaps, inconsistencies, problems, outliers, etc. in accordance with the Official Rules. Given that deaths are relatively unusual occurrence at almost all ages covered in the study, the data will display a significant amount of random fluctuation. Hence, we expect entrants not simply to flag all items that could be the result of anticipated random variations; thought needs to be given to isolate areas where the likelihood of error is greater than purely random variation.

Entrants should consider outlier, data validation and predictive analytics techniques to review the dataset, but always with a consideration as to whether the results obtained appear to be actual or potential errors or results that are readily explainable based on the fact that this is life insurance data.

Examples

This dataset contains some known problems that are examples of the types of data issues we are looking for the analyses to uncover. For example, the labeling of juvenile (<18 at issue) records with a "smoker" status has been identified as an issue.

Other potential "problems" are subtler, and knowledge of the insurance business would be beneficial. One example of what appears to be an error, but is not, is as follows. Consider the two-way table created from the data, with Insurance Plan being the columns and Duration the rows. The values for Insurance Plans Perm and Term only are shown below. If one calculates the Actual to Expected ratios (so taking the actual number of deaths in a category versus those expected, in this case using the VBT 2015 table as the expected basis using the expected values provided in the dataset), one would get the following ratios:

Duration	Perm	Term
1	2.30	1.44
2	1.90	1.29
3	1.68	1.18
4	1.51	1.08
5-9	1.49	1.01
10-14	1.37	1.01
15-19	1.22	1.13
20-25	1.16	1.43

Generally, we anticipate the actual to expected ratios to decrease as the duration increases. However, in the case of Term policies, this is not necessarily the case, and the above results are not an error. Many term policies have a level period of 15 years, and the policyholders who renew after 15 years are generally those whose health has deteriorated during the 15-year term and so are unable to get a better rate elsewhere. Hence the actual experience for the term policies gets worse after 15 years.