

The Business Problem

You are a data analyst for a California based insurer named NorCal Home Insurance Company (NCHIC). NCHIC is a writer of Homeowners insurance in Northern California and plans on applying predictive analytics to more accurately price its Homeowners product. As a result of the large wildfires in 2017, and the severe impact on NCHIC's book of business, the company is looking into new risk management solutions against such catastrophes.

At NCHIC, the Data Analytics department is responsible for modeling, and the Actuarial department is responsible for rate-making, reserving, product management and other traditional actuarial work. As a data analyst, you are tasked with the modeling aspect of this project and need to provide statistically sound rating variables and relativities to the actuarial team. Your team has access to two separate data tables: one with policy information and another with 7 years of claims history (NCHIC has had a constant customer base for the past 7 years). These data have been provided to you in two different formats: an Excel workbook and two separate CSV files.

Task 1 - Define Exposure

Research on "exposure base" and what exposure means in general, explain why it is important in terms of pricing for insurance. What is the most commonly used exposure base for homeowner insurance?

Remark: Here is an equation related to exposures that you would need in the following questions:
claim frequency = claim count/total exposure.

Task 2 - Identify Catastrophe

- (a) Identify any **major** catastrophes in the last 7 years. We suggest you explore the claim data by plotting the claim severity against time in R. Recall that claim severity = claim amount/claim count.
- (b) Explain how is catastrophe risk management different from traditional insurance risk management. Think about the feature of catastrophic claims, which can also see from the graph you produced above.
- (c) Before you start any calculations, explain what should be done to the catastrophic claims in the dataset and execute the corresponding actions that you just mentioned in R.

Hint: An R package that might be helpful here is **lubridate** when you are producing the graph.

Remark: This task is part of the exploratory data analysis and when you're trying to modify the data in part (c), think about how those catastrophic claims would affect the premium eventually.

Task 3 - Generalized Linear Model

Some Background: Generalized Linear Model is a technique of predictive modeling, and it gives the expected value of the response variable.

- (a) Explain why Poisson can be one option for the family distribution for frequency, and Gamma for the family distribution of severity when you use Generalized Linear Model.
- (b) Research on Tweedie distribution and give a brief description about it. Explain why Tweedie distribution describes pure premium.
- (c) Identify any correlated factors and the pricing factors that do not seem to affect claims significantly in R. What would you do to those not as related factors in the later analysis?
- (d) Exhibit 3 shows the base class (i.e. relativity of 1) determined by the actuarial department. Using this information and the 7 years of claims data, calculate a new set of relativities for each pricing factor you chose based on the analysis in part (b) using the information given in previous parts. To achieve this, calculate separate sets of relativities for frequency and severity, and then derive pure premium relativities.

Hint: For part (c), notice that you have some categorical variables and please think/research how to measure the correlations for them. Also, a package called **dummies** will be useful for processing categorical variables before running GLM.

Remark: To make life easier, **pure premium = frequency × severity**. For the midterm project we did not have the base class, but here you are given the base class for each rating factors. Therefore, the relativities for other classes are simply the ratios to the base class.

Task 4 - Calculate Premium and Give Business Solutions

Unfortunately catastrophes are unforeseen and the increasing frequency of California wildfires is urging the insurance companies to take actions.

- (a) With the relativities you calculated in Task 3, describe how you will calculate the premium to charge for each policy given the information in Exhibit 3. Just explain how you would do it in R, no calculations needed.
- (b) Given that catastrophes are so destructive that it is important for the insurance companies to get prepared for them, suggest two possible business solutions that the insurance companies can take to possibly reduce their loss resulting from these catastrophes.