1

Pure Premium Modeling Using Generalized Linear Models

Ernesto Schirmacher

Chapter Preview. Pricing insurance products is a complex endeavor that requires blending many different perspectives. Historical data must be properly analyzed, socioeconomic trends must be identified, and competitor actions and the company's own underwriting and claims strategy must be taken into account. Actuaries are well trained to contribute in all these areas and to provide the insights and recommendations necessary for the successful development and implementation of a pricing strategy. In this chapter, we illustrate the creation of one of the fundamental building blocks of a pricing project, namely, pure premiums. We base these pure premiums on generalized linear models of frequency and severity. We illustrate the model building cycle by going through all the phases: data characteristics, exploratory data analysis, one-way and multiway analyses, the fusion of frequency and severity into pure premiums, and validation of the models. The techniques that we illustrate are widely applicable, and we encourage the reader to actively participate via the exercises that are sprinkled throughout the text; after all, *data science is not a spectator sport*!

1.1 Introduction

The pricing of insurance products is a complex undertaking and a key determinant of the long-term success of a company. Today's actuaries play a pivotal role in analyzing historical data and interpreting socioeconomic trends to determine actuarially fair price indications.

These price indications form the backbone of the final prices that a company will charge its customers. Final pricing cannot be done by any one group. The final decision must blend many considerations, such as competitor actions, growth strategy, and consumer satisfaction. Therefore, actuaries, underwriters, marketers, distributors, claims adjusters, and company management must come together and collaborate on setting prices. This diverse audience must clearly understand price indications and the implications of various pricing decisions. Actuaries are well positioned to explain and

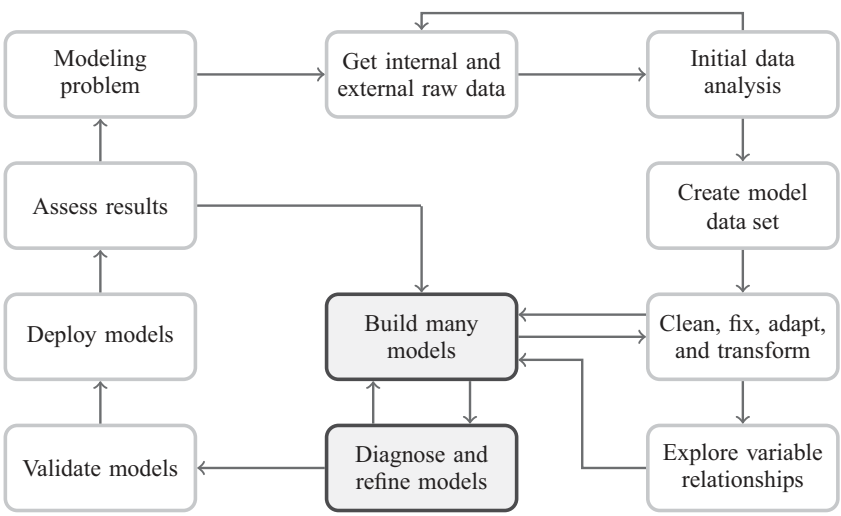

Fig. 1.1. Overall project cycle.

provide the insight necessary for the successful development and implementation of a pricing strategy.

Figure 1.1 shows one possible representation of an overall pricing project. Any one box in the diagram represents a significant portion of the overall project. In the following sections, we concentrate on the lower middle two boxes: "Build many models" and "Diagnose and refine models."

We concentrate on the first phase of the price indications that will form the key building block for later discussions, namely, the creation of pure premiums based on two generalized linear models. One model will address frequency, and the other one will target severity. These pure premiums are rooted in historical data.

Because our pure premiums only reflect historical data, they are unsuitable for use as price indications for a future exposure period. They lack the necessary trends (socioeconomic and company) to bring them up to date for the appropriate exposure period, and they also lack the necessary risk loadings, expense, and profit provisions to make them viable in the marketplace.

In Section 1.2, we describe the key overall characteristics of the dataset we have available, and in Section 1.3, we start exploring the variables. This dataset is an artificial private passenger automobile dataset that has many features that you will encounter with real datasets. It is imperative that you thoroughly familiarize yourself with the available data. The insights you gain as you explore the individual variables and their interrelationships will serve you well during the model construction phase.

In Sections 1.4 and 1.5, we start building models: frequency and severity, respectively. We illustrate several techniques that are widely applicable. We start with one-way analyses and move on to multiway analyses. The models we build are not necessarily the best ones, and we encourage the reader to explore and try to create better models. Data analysis is not a spectator sport. *The reader must actively participate*! To this end, we have sprinkled many exercises throughout the text. Most exercises require calculations that are best done in an environment that provides a rich set of data manipulation and statistical functions.

Exercise 1.1. Prepare your computing environment. Download the comma-delimited dataset `sim-modeling-dataset.csv` and load it into your environment.

All the exercises come with solutions (we have used the open-source `R` environment to illustrate the necessary calculations), but the reader will benefit most by looking at the solutions only after an honest attempt at their solution.

Section 1.6 combines the frequency and severity models to create pure premiums, and Section 1.7 shows some simple validation techniques on a portion of the data that our models have never seen. It is important that our modeling efforts do not overstate the accuracy or performance of the models we create. With today's available computing power and sophisticated algorithms, it is easy to overfit some models to the data. An overfit model tends to look very good, but when confronted with new data, its predictions are much worse.

Finally, Section 1.8 has some concluding remarks.

1.2 Data Characteristics

The modeling dataset `sim-modeling-dataset.csv` has been compiled for the actuarial department of a fictitious insurance company. The data set consists of private passenger automobile policy and claims information. It is an observational cross section of all in-force policies during the calendar years 2010 to 2013. There are a total of 40,760 rows and 23 variables[1] (see Table 1.1). The variables can be grouped into five classes: control variables, driver characteristics, geographic variables, vehicle characteristics, and response variables.

It is important to note that one record in our dataset can represent multiple claims from one insured. The variable `clm.count` measures the number of claims, and the variable `clm.incurred` has the sum of the individual claim payments and any provision for future payments; that is, it represents the ultimate settlement amount.

[1] The dataset actually contains 27 columns. One column identifies the rows (`row.id`). Driver age, years licensed, and vehicle age are represented by two columns each; one column is a string, and the other one is an integer.

Table 1.1. *Available Variables in Our Dataset*

Control	Driver	Vehicle	Geographic	Response
`year`	`age`	`body.code`	`region`	`clm.count`
`exposure`	`driver.gender`	`driver.age`		`clm.incurred`
`row.id`	`marital.status`	`vehicle.value`		
	`yrs.licensed`	`seats`		
	`ncd.level`	`ccm`		
	`nb.rb`	`hp`		
	`prior.claims`	`length`		
		`width`		
		`height`		
		`fuel.type`		

The variable `year` identifies the calendar year, and `exposure` measures the time a car was exposed to risk during the calendar year. We have one geographical variable, `region`, that tells us the garaging location of the vehicle. Unfortunately, the variable `region` has been coded as a positive integer, and we do not have any information about how these regions are spatially related. This is a significant drawback for our dataset and highlights that good data preparation is crucial. We want to retain as much information as possible.

The driver characteristic variables measure the age and gender of the principal operator of the vehicle, the marital status, the number of years that the operator has been licensed, the no claim discount level (higher level reflects a greater discount for not having any claims), and the number of prior claims. The variable `nb.rb` tells us whether this policy is new business (`nb`) or renewal business (`rb`).

The vehicle characteristic variables measure various attributes such as the body style (`body.code`); the age and value of the vehicle; the number of seats; and the vehicle's weight, length, width, and height. The variables `ccm`, `hp`, and `fuel.type` measure the size of the engine in cubic centimeters, the horsepower, and the type of fuel (gasoline, diesel, or liquefied petroleum gas), respectively.

We have two response variables, `clm.count` and `clm.incurred`, which measure the number of claims and the ultimate cost of those claims. All the variables in our dataset can be categorized as either continuous or categorical. Table 1.2 shows some summary statistics for the 14 continuous variables, and Table 1.3 has some information on the 12 categorical variables.

Overall frequency and severity statistics by calendar year across the entire dataset are given in Table 1.4. Note that the volume of business increased by about 78% from 2010 to 2012 and then decreased by 17% in 2013. Frequency for the first two years is at approximately 11% and then jumps up significantly to approximately 20%.

Table 1.2. *Summary Statistics for Continuous Variables*

Variable	Mean	Standard Deviation	Min.	Median	Max.
exposure	0.51	0.27	0.08	0.50	1.00
driver.age	44.55	10.78	18.00	44.00	93.00
yrs.licensed	3.21	1.89	1.00	3.00	10.00
vehicle.age	3.26	2.59	0.00	3.00	18.00
vehicle.value	23.50	8.89	4.50	22.10	132.60
ccm	1,670.69	390.12	970.00	1,560.00	3,198.00
hp	86.38	19.62	42.00	75.00	200.00
weight	1,364.22	222.01	860.00	1,320.00	2,275.00
length	4.32	0.36	1.80	4.28	6.95
width	1.78	0.10	1.48	1.74	2.12
height	1.81	0.09	1.42	1.82	2.52
prior.claims	0.83	1.33	0.00	0.00	21.00
clm.count	0.08	0.30	0.00	0.00	5.00
clm.incurred	66.52	406.23	0.00	0.00	11,683.58

Table 1.3. *Summary Statistics for Categorical Variables*

Variable	No. of Levels	Base Level	Most Common	Sample Levels
year	4	2013	2012	2010, 2011, 2012, 2013
nb.rb	2	NB	NB	NB, RB
drv.age	74	38	38	18, 19, 20, 21, 22, 23, 24, 25, and others
driver.gender	2	Male	Male	Female, Male
marital.status	4	Married	Married	Divorced, Married, Single, Widow
yrs.lic	8	1	2	1, 2, 3, 4, 5, 6, 7, 8+
ncd.level	6	1	1	1, 2, 3, 4, 5, 6
region	38	17	17	1, 10, 11, 12, 13, 14, 15, 16, and others
body.code	8	A	A	A, B, C, D, E, F, G, H
veh.age	15	1	1	0, 1, 10, 11, 12, 13, 14+, 2, and others
seats	5	5	5	2, 3, 4, 5, 6+
fuel.type	3	Diesel	Diesel	Diesel, Gasoline, LPG

Table 1.4. *Exposure, Claim Counts, Claim Amounts, Frequency, and Severity by Calendar Year for the Entire Dataset*

Year	Exposure	Claim Count	Claim Amount	Frequency	Severity
2010	3,661.9	422	287,869	0.115	682.2
2011	5,221.7	551	314,431	0.106	570.7
2012	6,527.2	1,278	1,021,152	0.196	799.0
2013	5,386.2	1,180	1,087,735	0.219	921.8
Total	20,797.1	3,431	2,711,187	0.165	790.2

The mean severity across all calendar years is at 790, but there are sharp increases over time, except for 2011, when we saw a decrease.

It is customary to split your data into three sets: training, testing, and validation. The training set is used to formulate your models. You do all the preliminary testing of your models against the testing set. The training and testing sets are used extensively to guide the development of your models and to try as best as possible to avoid both underfitting and overfitting. The validation set is used *only once* to determine how your final model will perform when presented with new data.

This three-way split of your data (train, test, validate) is only feasible when you have a large amount of data. In our case, we only have about 41,000 observations across four calendar years. This is a small dataset, so we will use a different testing and validation strategy, namely, cross-validation.

Rather than split our data into three sets, we will only split it into two sets: a training set and a validation set. We will use the training dataset to both develop and test our models. Because we only have one set of data for both training and testing, we cannot use standard testing techniques, so we will use k-fold cross-validation. We will set aside approximately 60% of our data as the training set.[2] The remainder will go in the validation set.

In k-fold cross-validation, we use all the training data to develop the structure of our models. Then, to test them, we split our training data into, say, five subsets called *folds*, and we label them 1 through 5. We set aside fold 1, combine folds 2 to 5, and estimate the parameters of our model on these data. Then we calculate our goodness-of-fit measure on fold 1 and set it aside. We repeat this procedure by setting aside fold 2, then fold 3, and so forth. At the end, we will have calculated five goodness-of-fit measures. We average them out, and that is our final goodness-of-fit estimate.

1.3 Exploratory Data Analysis

In this section, we start by exploring individual variables to gain a better understanding of the information we have available in our dataset. During exploratory data analysis, you want to concentrate on understanding how well each variable is populated, what kinds of values each variable takes, how missing values are coded, and the interrelationships between variables.

1.3.1 EDA for Frequency

From the previous section (see Table 1.4), we know that the overall frequency for the entire dataset is equal to 16.5%. For the training dataset, it is equal to 16.1%–very close to the overall frequency.

[2] We assigned a uniform random number, $u \in (0, 1)$, to each record. The training dataset consists of all records with $u < 0.6$, and the validation set consists of all those records with $u \geq 0.6$.

Table 1.5. *Frequency by Calendar Year and New/Renewal Business Indicator for the Training Dataset*

	Exposure		Claim Count		Frequency (%)	
Year	NB	RB	NB	RB	NB	RB
2010	1,504.5	684.6	193	67	12.8	9.8
2011	2,105.9	1,042.4	271	69	12.9	6.6
2012	2,643.5	1,278.5	551	189	20.8	14.8
2013	2,248.6	964.9	526	137	23.4	14.2
Total	8,502.5	3,970.4	1,541	462	18.1	11.6

Exercise 1.2. Add a random number u_i between 0 and 1 to each record. Calculate the frequency for all records with $u_i < 0.6$. How close is your estimate to the overall frequency of 16.5%? How variable is the frequency estimate as we resample the random numbers u_i?

We would like to understand how this frequency depends on the variables that we have at our disposal. Let's start by looking at the variable `nb.rb`. This variable is an indicator letting us know if the policy is new business (`NB`) or renewal business (`RB`). The frequency in our training dataset by this new/renewal business indicator is in Table 1.5. Notice that over the training data, the frequency for new business is equal to 18.1%, and for renewal business, it is equal to 11.6%. This looks like a significant difference; thus this variable is a good candidate to include in our models. Also note that on a year-by-year basis, there is a gap between the new and renewal business frequency. The gap for the last three years is quite large.

Exercise 1.3. What is the frequency of each region on the entire dataset? Has it been stable over time?

Next we can look at `driver.age`. This variable tells us the age of the principal operator of the vehicle. In the training data, we have ages 18 to 87, 89 to 90, and 93, for a total of 73 unique ages.[3]

Exercise 1.4. Verify that age 88 is not in the training dataset but that it is in the validation dataset. How should our modeling deal with such situations?

We should be suspicious of some of these very advanced ages and check that our data are accurate. Also, we should check how much exposure we have for all ages. A

[3] For the entire dataset, we have 74 unique ages. Age 88 is not represented in the training dataset but is in the validation dataset.

six-point summary over the training dataset for the frequency of claims by driver age[4] is

Min.	Q1	Q2	Mean	Q3	Max.
0%	9.2%	14.5%	16.8%	18.5%	184.6%

where Qn stands for the nth quartile. Note that the maximum frequency is equal to 184.6%, and upon looking into our data, we know it comes from four policies:

Row ID	Driver Age	Exposure	Claim Count
2885	19	1.000	1
2886	19	0.083	0
14896	19	0.167	0
14897	19	0.917	3

Also, the next highest frequency value is equal to 92.3%, and it comes from the two policies with drivers aged 89 years old that are in our training dataset. These two policies have a total exposure of 1.083 car-years and one claim.

Exercise 1.5. Check the exposure and number of claims for all the drivers aged 76 years old in the training dataset.

Figure 1.2 shows the driver age frequencies together with the amount of exposure. Clearly there is an overall decreasing frequency trend as driver age increases. The bulk of the exposure (approximately 98%) is concentrated in the age range from 25 to 70. Note that even though the frequency trend is decreasing, there is significant volatility in the individual driver age frequencies. For example, in the age range from 30 to 34 there is a zigzag pattern:

Driver age	30	31	32	33	34
Frequency	20.0%	18.2%	20.3%	22.4%	19.5%

Similar zigzag patterns occur between the ages of 50 to 70. Also there seems to be a spike in frequency around 47 years old. This could be due to young drivers using their parents' cars.

We have been looking at the frequencies in our training dataset for the calendar years 2010, 2011, 2012, and 2013 combined. We must also check that these patterns are consistent from one calendar year to the next. Each calendar year has less exposure

[4] First we calculated the frequency for each individual age, and then took the six-point summary across the 73 frequencies.

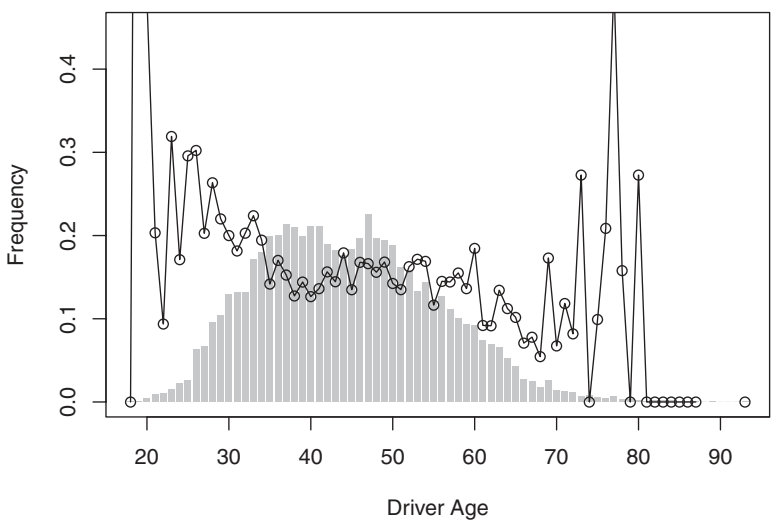

Fig. 1.2. Frequency and exposure by the driver age variable and for all three years of the training data. The y-axis has been restricted to the range [0, 0.45] to enhance the information shown. Four points have been omitted from the graph: (19, 184.6%), (20, 48.5%), (77, 50.3%), and (89, 92.3%).

than all three years combined, and so we expect that the individual calendar year frequency patterns will be more volatile.

Exercise 1.6. Create a graph similar to Figure 1.2, but add one frequency path for every calendar year.

Just as we have explored the frequency of claims by new or renewal business or by driver age, we should explore it across all other variables we have available in our dataset. We can mechanically create all sorts of tables and graphs for all variables at our disposal, but it would be better to concentrate our efforts on variables that we know from past experience have been important.

Exercise 1.7. Explore frequency by size of engine (variable `ccm`) for the entire dataset.

Exercise 1.8. Investigate the frequency of claims by the variables `driver.gender` and `marital.status`.

Exercise 1.9. From Exercise 1.8, we know that the frequency for married policyholders is about 15.8% and for widowers is about 27.3%. Is this difference significant? Is the difference in frequency between single and married policyholders significant?

Now let us shift attention to the variable `hp`. This variable represents the horsepower of the insured vehicle. In our training dataset, there are 63 unique values for horsepower ranging from a low of 42 to a high value of 200 but not all values between

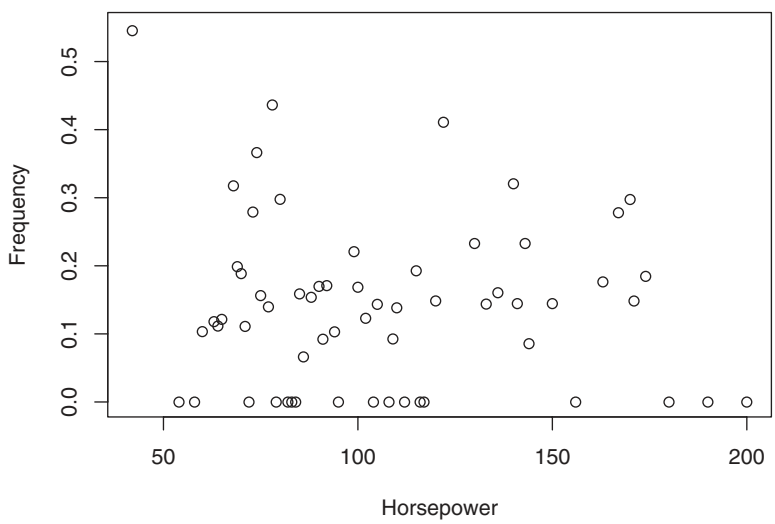

Fig. 1.3. Frequency by horsepower. To enhance the display of the data, the graph omits two frequency values: 0.585 and 0.923. The corresponding horsepower values are 48 and 125, respectively.

these two extremes are equally represented. The six largest frequencies by horsepower are 0.366, 0.411, 0.436, 0.545, 0.585, and 0.923. These six frequencies come from vehicles with horsepower equal to 74, 122, 78, 42, 48, and 125, respectively. It looks like the largest value might be an outlier. Also note that the exposure is concentrated in the following five values:

Horsepower	65	105	90	70	75
Exposure	848	1,442	1,631	2,486	2,628

These five values account for about 72% of the total exposure in the training dataset. Figure 1.3 does not show any systematic relationship between horsepower and frequency, so this variable is not a strong candidate for inclusion into the model for frequency that we develop in Section 1.4.

Another variable that is probably not a good predictor of frequency might be `length`; that is the length of the vehicle. In our dataset the unit of measurement for the variables `length`, `width`, and `height` is the meter. This unit is a bit awkward, so we transform these variables to use decimeters as the unit of measurement.

Exercise 1.10. Would we get different models if we use the variable `length` in units of meters or in units of decimeters?

Exercise 1.11. Explore the length of the vehicle variable. Start by describing the key characteristics of the lengths we have in our dataset. How many different lengths do we have? Are they uniformly distributed?