Ch 8. Risk classification

Contributing Author: Joseph H.T. Kim, *Yonsei University*

March 12, 2018

Chapter Preview. To be done.

1 Introduction

Through insurance contracts, the policyholders effectively transfer their risks to the insurer in exchange for premiums. For the insurer to stay in business, the premium income collected from a pool of policyholders must at least equal to the benefit outgo. Ignoring the frictional expenses associated with the administrative cost and the profit margin, the net premium thus should be equal to the expected loss occurring from the risk that is transferred from the policyholder to the insurer.

If all policyholders in the insurance pool have identical risk profiles, the insurer may simply charge the same premium for each policyholder because all of them would have the same expected loss. In reality however the policyholders are hardly homogeneous. For example, mortality risk in life insurance depends on the characteristics of the policyholder, such as, age, sex and life style. In auto insurance, those characteristics may include age, occupation, the type or use of the car, and the area where the driver resides. The knowledge of these characteristics or variables of individual policyholders can enhance the ability of calculating a fair premium as they can be used to estimate or predict the expected losses more accurately at the individual level. Indeed, if the insurer do not differentiate the risk characteristics of individual policyholders and simply charges the same premium to all individuals based on the average characteristic of the portfolio, the insurer would face adverse selection, a situation where individuals with a higher chance of loss are attracted in the portfolio and low-risk individuals are repelled.

For example, consider a health insurance industry where smoking status is an important risk factor for mortality and morbidity. Most health insurers in the industry require different premiums depending on smoking status, so smokers pay higher premiums than non-smokers, with other characteristics being identical. Now suppose that there is an insurer, we will call EquitabAll, that offers the same premium to all insureds regardless of smoking status, unlike other competitors. The net premium of EquitabAll is natually an average mortality loss accounting for both smokers and non-smokers; the average is a weighted one using the proportion of smokers and non-smokers. Thus it is easy to see that that a smoker would have a good incentive to purchase insurance from EquitabAll than other insurers as the offered premium by EquitabAll is relatively lower. At the same time non-smokers would

prefer buying insurance from somewhere else where lower premiums, computed from the non-smoker group only, are offered. The result of this tendency for the EquitabAll's insurance portfolio is that there will be more smokers and less non-smokers in the pool, which leads to larger-than-expected mortality losses and hence a higher premium for insureds in the next period to cover the higher costs. With the raised new premium in the next period, non-smokers in EquitabAll will have even greater incentives to switch the insurer. As the cycle continues over time, EquitabAll would gradually retain more smokers in its portfolio with the premium continually raised, and this vicious cycle eventually leads to a collapsing of business. In the literature this phenomenon is known as the *adverse selection spiral* or death spiral. Therefore, incorporating and differentiating important risk characteristics of individuals in the insurance pricing process are a pertinent component for both the determination of fair premium for each policyholder and the long term sustainability of an insurer.

In order to incorporate relevant risk characteristics of policyholders in the pricing process insurers maintain some classification system that assigns each policyholder to one of the risk classes based on a relatively small number of risk characteristics that are deemed most relevant. These characteristics used in the classification system are called the rating factors, which are *a priori* variables in the sense that they are known before the contract begins (e.g., sex, health status, vehicle type, etc, are known during the underwriting process). All policyholders sharing identical risk factors thus are assigned to the same risk class, and are considered homogeneous; the insurer consequently charge them the same premium.

An important task in any risk classification is to construct a quantitative model that can determine the expected loss given various rating factors for a policyholder. The standard approach is to adopt a statistical regression model which produces the expected loss as the output when the relevant risk factors are given as the inputs. We introduce and discuss the Poisson regression, which can be used when the loss is a count variable, as a prominent example of an insurance pricing tool under risk classification schemes.

2 Poisson regression model

The Poisson regression model has been successfully used in a wide range of applications and has an advantage of allowing closed-form expressions for important quantities, which provides a better intuition. In this section we introduce the Poisson regression as a natural extension of the Poisson distribution.

2.1 Need of Poisson regression

Poisson distribution

To introduce the Poisson regression, let us consider a hypothetical health insurance portfolio where all policyholders are of the same age and only one risk factor, smoking status, is relevant. Smoking status thus is a categorical variable containing two different types: smoker and non-smoker. In the statistical literature different types in a given categorical variable are commonly called *levels*. Thus we have two levels in our example, and may denote smoker and non-smoker by level 1 and 2, respectively[1]. Suppose now that we are interested in pricing a

[1]Here the numbering is arbitrary and nominal.

health insurance where the premium for each policyholder is determined by the number of outpatient visits to doctor's office during a year. The amount of medical cost for each visit is assumed to be the same regardless of the smoking status for simplicity. Thus if we believe that smoking status is a valid risk factor in this health insurance, it is natural to consider the data separately for each smoking status. In Table 1 we present the data for this portfolio. As this dataset contains random counts we try to fit a Poisson distribution for each level.

Table 1: Number of visits to doctor's office in last year

Smoker (level 1)		Non-smoker (level 2)		Both	
Count	Observed	Count	Observed	Count	Observed
0	2213	0	6671	0	8884
1	178	1	430	1	608
2	11	2	25	2	36
3	6	3	9	3	15
4	0	4	4	4	4
5	1	5	2	5	3
Total	2409	Total	7141	Total	9550
Mean	0.0926	Mean	0.0746	Mean	0.0792

The pmf of the Poisson with mean μ is given by

$$\Pr(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, \qquad y = 0, 1, 2, \ldots \tag{1}$$

and $\mathrm{E}(Y) = \mathrm{Var}(Y) = \mu$. Furthermore, the MLE of the Poisson distribution is given by the sample mean. Thus if we denote the Poisson mean parameter for each level by μ_1 (smoker) and μ_2 (non-smoker), we see from the table that $\hat{\mu}_1 = 0.0926$ and $\hat{\mu}_2 = 0.0746$. This simple example shows the basic idea of risk classification. Depending on the smoking status a policyholder will have a different risk characteristic and it can be incorporated through varying Poisson parameter in computing the fair premium. In this example the expected loss frequency ratio is $\hat{\mu}_1/\hat{\mu}_2 = 1.2402$, implying that smokers tend to visit doctor's office 1.2402 times more frequently compared to non-smokers.

It is also informative to note that if the insurer charges the same premium to all individuals based on the average characteristic of the portfolio, as is the case for EquitabAll described in Introduction, the premium $\hat{\mu}$ can be computed from the last column of Table 1, again using the Poisson MLE, as

$$\hat{\mu} = \left(\frac{n_1}{n_1 + n_2}\right)\hat{\mu}_1 + \left(\frac{n_2}{n_1 + n_2}\right)\hat{\mu}_2 = 0.0792, \tag{2}$$

where n_i is the number of observations in each level. Clearly, this premium is a weighted average of the premiums for each level with the weight equal to the proportion of the insureds in that level.

A simple Poisson regression

In the example above, we have fitted a Poisson distribution for each level separately, but we can actually combine them together in a unified fashion so that a single Poisson model can encompass both smoking and non-smoking statuses. This can be done by relating the Poisson mean parameter with the risk factor. In other words, we make the Poisson mean, which is the expected loss frequency, respond to the change in the smoking status. The conventional approach to deal with a categorical variable is to adopt indicator variables[2] that take either 1 or 0, so that we turn the switch on for one level and off for others. So we may propose to use

$$\mu = \beta_0 + \beta_1 x_1 \tag{3}$$

or, more commonly, a log linear form

$$\log \mu = \beta_0 + \beta_1 x_1, \tag{4}$$

where x_1 is an indicator variable with

$$x_1 = \begin{cases} 1 & \text{if smoker,} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

We generally prefer the log linear relation (4) to the linear one (3) to prevent undesirable events of yielding a negative μ value, which may happen when there are many different risk factors. The setup (4) and (5) then results in different Poisson frequency parameters depending on the level in the risk factor:

$$\log \mu = \begin{cases} \beta_0 + \beta_1 \\ \beta_0 \end{cases} \quad \text{or equivalently,} \quad \mu = \begin{cases} e^{\beta_0+\beta_1} & \text{if smoker (level 1),} \\ e^{\beta_0} & \text{if non-smoker (level 2),} \end{cases} \tag{6}$$

achieving what we intended to obtain. This is the simplest form of the Poisson regression. It is noted that we need a single indicator variable to model two levels. Alternatively, it is also possible to use two indicator variables through a different coding scheme. This scheme requires dropping the intercept term so that (4) is modified to

$$\log \mu = \beta_1 x_1 + \beta_2 x_2, \tag{7}$$

where x_2 is the second indicator variable with

$$x_2 = \begin{cases} 1 & \text{if non-smoker,} \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Then we have, from (7),

$$\log \mu = \begin{cases} \beta_1 \\ \beta_2 \end{cases} \quad \text{or} \quad \mu = \begin{cases} e^{\beta_1} & \text{if smoker (level 1),} \\ e^{\beta_2} & \text{if non-smoker (level 2).} \end{cases} \tag{9}$$

[2] Also known as binary or dummy variables.

The numerical result of (6) is the same as (9) as all the coefficients are given as numbers in actual applications, with the former setup more common in most texts; we also stick to the former.

With this Poisson regression we can easily show how the coefficients β_0 and β_1 are linked to the expected loss frequency in each level. According to (4), the Poisson mean parameter of the smokers, μ_1, is given by $\mu_1 = e^{\beta_0} e^{\beta_1} = \mu_2\, e^{\beta_1}$ where μ_2 is the Poisson mean for the non-smokers. This relation between the smokers and non-smokers suggests a useful way to compare the risks embedded in different levels of a given risk factor. That is, the increase in the loss frequency of the smokers compared to that of the non-smokers is modelled by a multiplicative factor e^{β_1}, provided that $\beta_1 > 0$. Putting another way, if we set the expected loss frequency of the non-smokers as the base value, the expected loss frequency of the smokers is obtained by applying e^{β_1} to the base value.

Dealing with multi-level case

We can readily extend the two-level case to a multi-level one by considering l different levels for a single rating factor. For this we generally need $l-1$ indicator variables to formulate

$$\log \mu = \beta_0 + \beta_1 x_1 + \ldots + \beta_{l-1} x_{l-1}, \tag{10}$$

where x_k is an indicator variable that takes 1 if the policy belongs to level k and 0 otherwise, for $k = 1, 2, \ldots, l-1$. By omitting the indicator variable associated with the last level in the formulation (10) we effectively chose level l as the base case, but this choice is arbitrary and does not matter numerically. The resulting Poisson mean parameter for policies in level i then becomes, from (10),

$$\mu = \begin{cases} e^{\beta_0+\beta_k} & \text{if the policy belongs to level } k \ (k = 1, 2, \ldots, l-1), \\ e^{\beta_0} & \text{if the policy belongs to level } l. \end{cases} \tag{11}$$

Thus if we denote the Poisson parameter for policies in level k by μ_k, we can relate the Poisson parameter for different levels through $\mu_k = \mu_l\, e^{\beta_k}$, $k = 1, 2, \ldots, l-1$, indicating that the expected loss frequency of the ith level is that of the base value multiplied by the relative factor e^{β_k}. This relative interpretation becomes more powerful as there are more risk factors with many levels in the model, and leads us to a better understanding of the underlying risk and more accurate prediction of future losses. Finally, we note that the varying Poisson mean is completely driven by the coefficient parameters β_k's, which are to be estimated from the dataset; the procedure of the parameter estimation will be explained later in this chapter.

2.2 Poisson regression

We now describe the Poisson regression in a formal and more general setting. Let us assume that there are n independent policyholders with a set of rating factors characterised by a k-variate vector[3]. The ith policyholder's rating factor is thus denoted by vector $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik})'$, and the policyholder has recorded the loss count $y_i \in \{0, 1, 2, \ldots\}$, for

[3]For example, if there are 3 risk factors with the number of levels 2, 3 and 4, respectively, $k = (2-1) \times (3-1) \times (4-1) = 6$.

$i = 1, \ldots, n$. In the regression literature, the values $x_{i1}, \ldots, x_{ik}$ are generally known as the *explanatory variables*, as these are measurements providing information about the variable of interest y_i. In essence, regression analysis is a method to quantify the relationship between a variable of interest and explanatory variables.

We also assume, for now, that all policyholders have the same one unit period for loss observation, or equal exposure[4] of 1, to keep things simple. As previously mentioned, we now describe the Poisson regression through its mean function. For this we first denote μ_i to be the expected loss count of the ith policyholder under the Poisson specification (1):

$$\mu_i = \mathrm{E}(y_i|\mathbf{x}_i), \qquad y_i \sim Pois(\mu_i),\ i = 1, \ldots, n. \tag{12}$$

Note that the subscript for μ was used for different levels in Section 2.1, but here it represents individual policyholders. The condition inside the expectation operation in (12) indicates that the loss frequency is the model output responding to the given set of risk factors or explanatory variables. In principle the conditional mean $\mathrm{E}(y_i|\mathbf{x}_i)$ in (12) can take different forms depending on how we specify the relationship between $\mathbf{x}$ and y. The standard choice for the Poisson regression is to adopt the exponential function, as we mentioned previously, so that

$$\mu_i = \mathrm{E}(y_i|\mathbf{x}_i) = e^{\mathbf{x}_i'\beta}, \qquad y_i \sim Pois(\mu_i),\ i = 1, \ldots, n. \tag{13}$$

Here $\beta = (\beta_0, \ldots, \beta_k)'$ is the vector of coefficients so that $\mathbf{x}_i'\beta = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$. The exponential function in (13) ensures that $\mu_i > 0$ for any set of rating factors $\mathbf{x}_i$ and each y_i appropriately forms count data generated from the Poisson, a non-negative discrete distribution. Often (13) is rewritten as

$$\log \mu_i = \log \mathrm{E}(y_i|\mathbf{x}_i) = \mathbf{x}_i'\beta, \qquad y_i \sim Pois(\mu_i),\ i = 1, \ldots, n \tag{14}$$

to reveal the relationship when the right side is set as the linear form, $\mathbf{x}_i\beta$. Again, we see that the mapping works well as both sides of (14), $\log \mu_i$ and $\mathbf{x}_i\beta$, can cover the entire real values. This is the formulation of the Poisson regression, assuming that all policyholders have the same unit period of exposure. When the exposures differ among the policyholders, however, as is the case in most practical cases, we need to revise this formulation by adding exposure component.

2.3 Incorporating exposure

The assumption of the same unit of exposure across all policyholders is not a realistic one. In auto insurance, for example, two different drivers with different lengths of insurance coverage (e.g., 3 months and 12 months, respectively) could have recorded the same number of accidents. As the expected number of accidents would be proportional to the length of coverage, we should not treat these two drivers' loss experiences identically in the modelling process. This motivates the need of the concept of the *exposure* in the Poisson regression.

The Poisson distribution in (1) is parametrised via its mean. To understand the exposure, we alternatively parametrize the Poisson pmf in terms of the *rate* parameter λ, inspired from

[4] We will discuss more details on the exposure in the following subsection.

the definition of the Poisson process:

$$\Pr(Y = y) = \frac{(\lambda t)^y e^{-\lambda t}}{y!}, \qquad y = 0, 1, 2, \ldots \tag{15}$$

with $\mathrm{E}(Y) = \mathrm{Var}(Y) = \lambda t$. Here λ is understood as the rate or intensity per unit period of the Poisson process and t represents the length of time or *exposure*, a known constant. For given λ the Poisson distribution (15) produces a larger expected loss count as the exposure t gets larger. Clearly, (15) reduces to (1) when $t = 1$, which means that the mean and the rate become the same for the unit exposure, the case we considered in the previous subsection. In principle the exposure does not need to be measured in units of time and may represent different things depending the problem at hand. For example,

1. In health insurance, the rate may be the occurrence of a specific disease per 1,000 people and the exposure is the number of people considered in the unit of 1,000.

2. In auto insurance, the rate may be the number of accidents per year of a driver and the exposure is the length of the observed period for the driver in the unit of year.

3. In marketing, the rate may be the number of customers who enter a store per hour and the exposure is the number of hours observed.

4. In civil engineering, the rate may be the number of major cracks on the paved road per 10 kms and the exposure is the length of road considered in the unit of 10 kms.

5. In biology, the rate may be the number of a specific type of bacteria found per 1 cm^3 of sea water and the exposure is the volume of sea water considered in the unit of cubic centimeter.

6. In credit risk modelling, the rate may be the number of default events per 1000 firms and the exposure is the number of firms under consideration in the unit of 1,000.

Therefore incorporating the exposure in the Poisson regression requires us to carefully separate the rate and exposure in the modelling process. Focusing on the insurance context, if we denote the rate of the loss event of the ith policyholder by λ_i, the known exposure (the length of coverage) by m_i and the expected loss count under the given exposure by μ_i, then the Poisson regression formulation in (13) and (14) should be revised in light of (15) as

$$\mu_i = \mathrm{E}(y_i|\mathbf{x}_i) = m_i\,\lambda_i = m_i\,e^{\mathbf{x}_i'\beta}, \qquad y_i \sim Pois(\mu_i),\ i = 1, \ldots, n, \tag{16}$$

and

$$\log \mu_i = \log m_i + \mathbf{x}_i'\beta, \qquad y_i \sim Pois(\mu_i),\ i = 1, \ldots, n. \tag{17}$$

Adding $\log m_i$ in (17) does not pose a problem in fitting as we can always specify this as an extra explanatory variable, as it is a known constant, and fix its coefficient to 1. In the literature $\log m_i$ is commonly called the *offset*.

2.4 Estimating Poisson regression model

Maximum likelihood estimation

In the Poisson regression the varying Poisson mean is determined by parameters β_i's, as shown in (17). In this subsection we use the maximum likelihood method to estimate these parameters. From (16) and (17), the log-likelihood for n observations is given by

$$\begin{aligned}\log L(\beta) = l(\beta) &= \sum_{i=1}^{n} \left(-\mu_i + y_i \log \mu_i - \log y_i!\right) \\ &= \sum_{i=1}^{n} \left(-m_i \exp(\mathbf{x}_i'\beta) + y_i \left(\log m_i + \mathbf{x}_i'\beta\right) - \log y_i!\right) \end{aligned} \tag{18}$$

To obtain the MLE of β, we differentiate[5] $l(\beta)$ with respect to vector β and set it to zero:

$$\frac{\partial}{\partial \beta} l(\beta)\bigg|_{\beta=\mathbf{b}} = \sum_{i=1}^{n} \left(y_i - m_i \exp(\mathbf{x}_i'\mathbf{b})\right) \mathbf{x}_i = \mathbf{0}. \tag{19}$$

Numerically solving this equation system gives the MLE of β, denoted by $\mathbf{b} = (b_0, b_1, \ldots, b_k)'$. Note that, as $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik})'$ is a column vector, Equation (19) is a system of $k+1$ equations with both sides written as column vectors of size $k+1$. If we denote $\hat{\mu}_i = m_i \exp(\mathbf{x}_i'\mathbf{b})$, we can rewrite (19) as

$$\sum_{i=1}^{n} \left(y_i - \hat{\mu}_i\right) \mathbf{x}_i = \mathbf{0}. \tag{20}$$

Since the solution $\mathbf{b}$ satisfies (20), it follows that the first among the array of $k+1$ equations, corresponding to the first constant element of $\mathbf{x}_i$, yields

$$\sum_{i=1}^{n} \left(y_i - \hat{\mu}_i\right) \times 1 = 0, \tag{21}$$

which implies that we must have

$$n^{-1} \sum_{i=1}^{n} y_i = \bar{y} = n^{-1} \sum_{i=1}^{n} \hat{\mu}_i. \tag{22}$$

This is an interesting property saying that the average of the individual losses, $\bar{y}$, is same as the average of the estimated values. That is, the mean is preserved under the fitted Poisson regression model.

Information matrix

Taking second derivatives gives the information matrix of the MLE estimators,

$$\mathbf{I}(\beta) = -\mathrm{E}\left(\frac{\partial^2}{\partial \beta \partial \beta'} l(\beta)\right) = \sum_{i=1}^{n} \mu_i \mathbf{x}_i \mathbf{x}_i'. \tag{23}$$

In practice, μ_i in (23) is replaced with $\hat{\mu}_i = m_i \exp(\mathbf{x}_i'\mathbf{b})$ to estimate the relevant variances and covariances of the MLE $\mathbf{b}$ or its functions.

[5] We use matrix derivative here.

2.5 Exercises

1. Regarding Table 1 answer the followings.
   (a) Verify the mean values in the table.
   (b) Verify the number in Equation (2).
   (c) Produce the fitted Poisson counts for each smoking status in the table.
2. In the Poisson regression formulation (12), consider using $\mu_i = \mathrm{E}(y_i|\mathbf{x}_i) = (\mathbf{x}_i'\beta)^2$, for $i = 1, \ldots, n$, instead of the exponential function What potential issue would you have?
3. Verify Equation (23) by differentiating the log-likelihood (18).

3 Categorical variables and multiplicative tariff

3.1 Rating factors and tariff

So far we have implicitly assumed that the risk characteristics of individual policyholders can be considerably diverse. However, this may not be so because in practice most rating factors in insurance are categorical variables, meaning that they take one of the pre-determined number of possible values. Examples of categorical variables include sex, type of cars, the driver's region of residence and occupation. Continuous variables, such as age or auto mileage, can also be grouped by bands and treated as categorical variables. Thus for a large insurer with only a few rating factors, there will be many policyholders in each risk class. For the remaining of this chapter we assume that all rating factors are categorical variables.

To further elaborate how rating factors are used, we consider a hypothetical auto insurer with only two rating factors:

- Type of vehicle: Type A (personally owned) and B (owned by corporations). We use index $j = 1$ and 2 for this rating variable.
- Age band of the main driver: Young (age < 25), middle ($25 \leq$ age < 60) and old age (age ≥ 60). We use index $k = 1, 2$ and 3, respectively, for this rating variable.

Thus there are $2 \times 3 = 6$ different risk classes in total for this insurer. From these classification rule, we may create an organized table or list called *tariff*, such as the one shown in Table 2, collected from all policyholders. These six different risk classes are called tariff cells. Each tariff cell represents a unique combination of the rating factors, and all policyholders in the same tariff cell are considered to be homogeneous and charged the same premium.

In the table the exposure means the sum of the length of insurance coverages, or in-force times, in the unit of year, of all the policyholders in each tariff cell. Similarly each claim count in the table is the number of claims at each cell. Clearly the exposures and claim counts vary due to the different number of drivers across the cells, as well as different in-force time periods among the drivers within each cell.

Table 2: Tariff of an illustrative auto insurer

Rating factors		Risk class (i)	Exposure	Claim count
Type (j)	Age (k)	(Tariff cell)	in year	observed
$j=1$	$k=1$	$i=1$	89.1	9
1	2	2	208.5	8
1	3	3	155.2	6
2	1	4	19.3	1
2	2	5	360.4	13
2	3	6	276.7	6

Now, we denote the exposure and claim count of cell (j,k) as m_{jk} and y_{jk}, respectively, and define the claim count per unit exposure as

$$z_{jk} = \frac{y_{jk}}{m_{jk}}, \qquad j = 1,2;\ k = 1,2,3. \tag{24}$$

For example, $z_{12} = 8/208.5 = 0.03837$, meaning that a policyholder in tariff cell (1,2) would have 0.03837 accidents if insured for a full year on average. The set of z_{ij} values then corresponds to the rate parameter in the Poisson distribution (15) as they are the event occurrence rates per unit exposure in the Poisson precess. We also note that the index pair (j,k) can be replaced by a single index i by mapping two risk factors to risk class as shown in the third column of the table, which sometimes can be convenient in our discussion. If there are many risk factors, the number of cells grows exponentially and the tariff would consist of a set of tables, instead of one, separated by some of the basic rating factors, such as sex or territory.

In order to apply the Poisson regression model for the tariff in Table 2, we need to convert the categorial rating factors into numerical values. This is the topic of the next subsection.

3.2 Multiplicative tariff model

We will assume the same illustrative auto insurer with two rating factors, car type and age. Recall that the loss count or frequency of a policyholder is described by the Poisson regression model with rate λ per unit exposure and the amount of exposure m, so that the expected loss count becomes $m\lambda$. As λ should respond to the change of the rating factors, we need to relate it to the rating factors in some functional form. Among others, we consider the following multiplicative functional relation, which is common in practice,

$$\lambda_{jk} = f_0 \times f_{1j} \times f_{2k}, \qquad j = 1,2;\ k = 1,2,3. \tag{25}$$

Here $\{f_{1j}, j = 1,2\}$ are the parameters associated with the two levels in the first rating factor, car type, and $\{f_{2k}, k = 1,2,3\}$ are associated with the three levels in the age band, the second rating factor. f_0 is some base value to be discussed later. Thus these parameters are the numerical counterparts of the categorial rating factors, and are to be estimated

from the dataset. The model (25) explains how the expected loss count (per unit exposure) changes as each rating factor varies. For example, if $f_{11} = 1$ and $f_{12} = 1.2$, then the expected loss count of a policyholder with vehicle type B would be 20% larger than the one with type B, when other factors are fixed. However, (25) has an identification issue because, for any $c > 0$, we have

$$\lambda_{ij} = f_0 \times \frac{f_{1j}}{c} \times c\, f_{2k}, \tag{26}$$

indicating that the rate parameter can be identical for very different individual rating factors. This over-parametrization, meaning that many different sets of parameters arrive at the identical model, obviously calls for some restriction on the rating factors. The standard practice is to make one parameter in each rating factor equal to one. This can be made arbitrarily, so we assume that $f_{11} = 1$ and $f_{21} = 1$ for our illustration. This way all other parameters are uniquely determined. The tariff cell $(1, 1)$ is then called the base cell, in which the rate is simply $\lambda_{11} = f_0$, the base value, according to (25). Thus we generally interpret f_0 as the base value that is equal to the Poisson rate of the base cell. In non-life insurance, the parameters f_{1j} and f_{2k} are known as price relatives as they determine how much expected loss, the net premium, changes relative to the base value f_0. Often (25) is log-transformed and rewritten as

$$\log \lambda_{jk} = \log f_0 + \log f_{1j} + \log f_{2k}, \tag{27}$$

as it is easier to work with a linear form in estimating process. It is also advantageous as log intensity in the left side can take any real value with $\lambda_{jk} > 0$, similar to (14).

The log-linear model (27) resembles the two-way analysis of variance (ANOVA) in Statistics, which studies how two different factors or treatments affect the mean outcome of quantity of interest. Thus we may adopt the standard statistical techniques for our tariff analysis.

3.3 Poisson regression for multiplicative tariff

We now turn to the estimation of these price relatives. For this we need to know how categorial variables can be incorporated in regression models. When a categorial variable is used as an explanatory variables in a regression model, it is customary to use an indicator variable which takes either 0 or 1. For our illustrative insurer, we may define for the first rating variable as

$$x_1 = \begin{cases} 1 & \text{for vehicle type B,} \\ 0 & \text{for vehicle type A.} \end{cases} \tag{28}$$

For the second rating variable, however, an indicator variable does not work directly as there are three possible levels. The trick is to use two indicator variables for the age band[6], that is,

$$x_2 = \begin{cases} 1 & \text{for age band 2,} \\ 0 & \text{otherwise.} \end{cases} \tag{29}$$

[6]If there are l levels for a given rating factor we need $l - 1$ indicator variables.

and

$$x_3 = \begin{cases} 1 & \text{for age band 3,} \\ 0 & \text{otherwise.} \end{cases} \tag{30}$$

The triple (x_1, x_2, x_3) then can effectively and uniquely determine each risk class. For example, cell $(j, k) = (2, 2)$ with vehicle type B and age band 2, corresponds to $(1, 1, 0)$, and the base cell $(j, k) = (1, 1)$ corresponds to $(0, 0, 0)$. Now we can express the log intensity (27) in terms of indicator variables as

$$\log \lambda_{jk} = \log f_0 + \log f_{12} \times x_1 + \log f_{22} \times x_2 + \log f_{23} \times x_3. \tag{31}$$

For example, for a policyholder at cell $(j, k) = (2, 2)$ we have $(x_1, x_2, x_3) = (1, 1, 0)$ to yield, from (31), $\log \lambda_{22} = \log f_0 + \log f_{22}$, and for the base cell $(j, k) = (1, 1)$ corresponding to $(x_1, x_2, x_3) = (0, 0, 0)$, we obtain $\log \lambda_{11} = \log f_0$.

It is important to note that we have carefully selected the coefficients for each indicator variable in (31), so that the base cell produces a vector of zeros, $(x_1, x_2, x_3) = (0, 0, 0)$. This way, the log intensity for the base cell yields $\lambda_{11} = f_0$ as required. Also note that we do not need to include f_{11} and f_{21} in (31) as $\log f_{11} = \log f_{21} = 0$ by construction. If we reverse the binary switch for x_i in (28) – (30) or select different coefficients in (31), we will lose this internal consistency; in an exercise question, the reader is invited to check this.

Under this specification, let us convert the tariff cell index (j, k) to a single risk class index i as shown in Table 2. Then for the ith risk class with triple (x_{i1}, x_{i2}, x_{i3}), we now recognize (31) as

$$\log \lambda_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} = \mathbf{x}_i' \beta, \qquad i = 1, \ldots, 6, \tag{32}$$

where $\beta_0, \ldots, \beta_3$ can be found in the corresponding position in (31). This in fact is exactly what we did in (14). Therefore, by incorporating the exposure in each risk class we can readily estimate parameters $\beta_0, \ldots, \beta_3$ and thus f_{1j} and f_{2k} in the same way as we did in (16) and (17). In Table 3 we present the mapping between β_i coefficients and the price relatives, keeping the same list form as Table 2. All columns except for the first is indexed with i, giving us a complete specification for our Poisson regression.

Table 3: Tariff parametrization of an illustrative auto insurer

Rating factors (j, k)	Risk class i	Covariates (x_{i1}, x_{i2}, x_{i3})	Price relativity $\lambda_i = e^{\mathbf{x}_i \beta}$	Exposure m_i	Claim count y_i
$(1, 1)$	1	$(0, 0, 0)$	$\lambda_1 = f_0 f_{11} f_{21}$	89.1	9
$(1, 2)$	2	$(0, 1, 0)$	$\lambda_2 = f_0 f_{11} f_{22}$	208.5	8
$(1, 3)$	3	$(0, 0, 1)$	$\lambda_3 = f_0 f_{11} f_{23}$	155.2	6
$(2, 1)$	4	$(1, 0, 0)$	$\lambda_4 = f_0 f_{12} f_{21}$	19.3	1
$(2, 2)$	5	$(1, 1, 0)$	$\lambda_5 = f_0 f_{12} f_{22}$	360.4	13
$(2, 3)$	6	$(1, 0, 1)$	$\lambda_6 = f_0 f_{12} f_{23}$	276.7	6

We can generalise this to a case where there many rating factors. Assume that there are p predictors, $(x_1, \ldots, x_p)$, after accounting for all relevant indicator variables. Then

, and each rating factor has s_k levels.

To do—

- table like Table 2.5 and 2.6 of Ohlsson and Johansson book?
- Generalise to many risk factor case. say s

4 Further Reading and References

It can be the number of people when μ is the

As most insurers maintain a risk classification system, individual policyholders are classified into one of the risk classes based on their rating factors. Consequently each risk class has a different number of policyholders and all policyholders in the same risk class would have the identical rating factors $\mathbf{x}$ and yield the same expected loss count from the Poisson regression model. This motivates the need of the concept of the exposure. For example, suppose that risk class 1 has 100 policyholders whereas risk class 2 has 10,000.

Therefore, we need It is common that the insurer the loss datasets typically At the beginning of this chapter we assume that i stands for Recall from the early chapters that when X_1 and X_2 are Poisson distributed independently with mean λ_1 and λ_2, respectively, the sum $X_1 + X_2$ is again Poisson distributed with mean $\lambda_1 + \lambda_1$. As an insurance example, if the ith risk class

denote μ_i as the expected value

In the Poisson regression, we use

$$\log \mu_i = \mathbf{x}_i \beta \tag{33}$$

or

$$\mu_i = e^{\mathbf{x}_i \beta} \tag{34}$$

- $$y|\mathbf{x} = f(\mathbf{x}, \beta) + \varepsilon \tag{35}$$

  for some linear or non-linear function $f$with $E(\varepsilon) = 0$
  (Note that we us lower case y and $\mathbf{x}$ for GLM studies, as this could be more convenient)

- Equivalent to the above, the mean response is given by

$$E(y|\mathbf{x}) = f(\mathbf{x}, \beta) := \mu \tag{36}$$

- As we like $E(y|\mathbf{x})$ to vary by each level of the explanatory variable, we denote the mean response for each observation as

$$\mu_i = E(y_i|\mathbf{x}_i) = f(\mathbf{x}_i, \beta) \tag{37}$$

$$\tag{38}$$

  where $\mathbf{x}_i = (x_{i1}, \cdots, x_{ik})$ and $\beta = (\beta_1, \cdots, \beta_k)$.

5 Application

6