Ch 8. Risk classification

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

May 7, 2018

Chapter Preview. This chapter motivates the use of risk classification in insurance pricing and introduces readers to the Poisson regression as a prominent example of risk classification. In Section 1 we explain why insurers need to incorporate various risk characteristics, or rating factors, of individual policyholders in pricing insurance contracts. We then introduce the Poisson regression as a pricing tool to achieve such premium differentials. The concept of exposure is also introduced in this section. As most rating factors are categorical, we show in Section 3 how the multiplicative tariff model can be incorporated in the Poisson regression model in practice, along with numerical examples for illustration.

1 Introduction

In this section you learn:

- Why premiums should vary across policyholders with different risk characteristics.
- The meaning of the adverse selection spiral.
- The need of risk classification.

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 charged by the insurer thus should be equal to the expected loss occurring from the risk that is transferred from the policyholder.

If all policyholders in the insurance pool have identical risk profiles, the insurer simply charges the same premium for all policyholders because they 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 can enhance the ability of calculating fair premiums for individual policyholders as they can be used to estimate or predict the expected losses more accurately.

Indeed, if the insurer do not differentiate the risk characteristics of individual policyholders and simply charges the same premium to all insureds based on the average loss in 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 market 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. That is, the net premium is a weighted average of the losses with the weights being the proportion of smokers and non-smokers, respectively. Thus it is easy to see that that a smoker would have a good incentive to purchase insurance from EquitabAll than from 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. As a result, there will be more smokers and less non-smokers in the EquitabAll's portfolio, which leads to larger-than-expected 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 this cycle continues over time, EquitabAll would gradually retain more smokers and less non-smokers in its portfolio with the premium continually raised, eventually leading 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 individual policyholders and the long term sustainability of insurers.

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). All policyholders sharing identical risk factors thus are assigned to the same risk class, and are considered homogeneous from the pricing viewpoint; 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 of 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. In this chapter we learn the Poisson regression, which can be used when the loss is a count variable, as a prominent

example of an insurance pricing tool.

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 informative intuition and interpretation. In this section we introduce the Poisson regression as a natural extension of the Poisson distribution.

In this section you will:

- Understand Poisson regressions as convenient tool to combine individual Poisson distributions in a unified fashion.
- Learn the concept of exposure and its importance.
- Formally learn how to formulate the Poisson regression model using indicator variables when the explanatory variables are categorical.

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*. As there are two levels for the smoking status, we may denote smoker and non-smoker by level 1 and 2, respectively. Here the numbering is arbitrary and nominal. Suppose now that we are interested in pricing a 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.

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 $\mu_{(1)}$ (smoker) and $\mu_{(2)}$ (non-smoker), we see from Table 1 that $\hat{\mu}_{(1)} = 0.0926$ and $\hat{\mu}_{(2)} = 0.0746$. This

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

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 ratio of expected loss frequencies is $\hat{\mu}_{(1)}/\hat{\mu}_{(2)} = 1.2402$, implying that smokers tend to visit doctor's office 24.02% times more frequently compared to non-smokers.

It is also informative to note that if the insurer charges the same premium to all policyholders regardless of the smoking status, based on the average characteristic of the portfolio, as was the case for EquitabAll described in Introduction, the expected frequency (or the premium) $\hat{\mu}$ is 0.0792, obtained from the last column of Table 1. It is easily verified that

$$\hat{\mu} = \left(\frac{n_1}{n_1+n_2}\right)\hat{\mu}_{(1)} + \left(\frac{n_2}{n_1+n_2}\right)\hat{\mu}_{(2)} = 0.0792, \tag{2}$$

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 or dummy variables that take either 1 or 0, so that we turn the switch on for one level and off for others. Therefore 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 in (3) to prevent undesirable events of producing negative μ values, which may happen when there are many different risk factors and levels. 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 aim for. This is the simplest form of the Poisson regression. Note that we require a single indicator variable to model two levels in this case. 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),

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

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

With this Poisson regression model we can easily understand how the coefficients β_0 and β_1 are linked to the expected loss frequency in each level. According to (6), the Poisson mean of the smokers, $\mu_{(1)}$, is given by

$$\mu_{(1)} = e^{\beta_0+\beta_1} = \mu_{(2)}\, e^{\beta_1} \quad \text{or} \quad \mu_{(1)}/\mu_{(2)} = e^{\beta_1} \tag{10}$$

where $\mu_{(2)}$ 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 proportional increase in the expected loss frequency of the smokers compared to that of the non-smokers is simply given by a multiplicative factor e^{β_1}. 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 where l different levels are involved for a single rating factor. For this we generally need $l-1$ indicator variables to formulate

$$\log \mu = \beta_0 + \beta_1 x_1 + \ldots + \beta_{l-1} x_{l-1}, \tag{11}$$

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 (11) we effectively chose level l as the base case, but this choice is arbitrary and does not matter numerically. The resulting Poisson parameter for policies in level k then becomes, from (11),

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

Thus if we denote the Poisson parameter for policies in level k by $\mu_{(k)}$, we can relate the Poisson parameter for different levels through $\mu_{(k)} = \mu_{(l)}\, e^{\beta_k}$, $k = 1, 2, \ldots, l-1$. This indicates that, just like the two-level case, the expected loss frequency of the kth level is obtained from the base value multiplied by the relative factor e^{β_k}. This relative interpretation becomes more powerful when there are many risk factors with multi-levels, 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 discussed 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 characterized by a k-variate vector[1]. 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 $y_i \in \{0, 1, 2, \ldots\}$ from the last period of loss observation, for $i = 1, \ldots, n$. 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 of 1, to keep things simple; we will discuss more details on the exposure in the following subsection. As done before, we 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):

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

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

The condition inside the expectation operation in (12) indicates that the loss frequency μ_i 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 $\mathbf{x}_i$. Often (13) is rewritten as a log linear form

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

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 now 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 as an additional term in (14).

2.3 Incorporating exposure

Concept of exposure

In order to determine the size of potential losses in any type of insurance, one must always know the corresponding exposure. The concept of exposure is an extremely important ingredient in insurance pricing, though we usually take it for granted. For example, when we say the expected claim frequency of a health insurance policy is 0.2, it does not mean much without the specification of the exposure such as, in this case, per month or per year. In fact, all premiums and losses need the exposure precisely specified and must be quoted accordingly; otherwise all subsequent statistical analyses and predictions will be distorted.

In the previous section we assumed the same unit of exposure across all policyholders, but this is hardly realistic in practice. In health insurance, for example, two different policyholders with different lengths of insurance coverage (e.g., 3 months and 12 months, respectively) could have recorded the same number of claim counts. As the expected number of claim counts would be proportional to the length of coverage, we should not treat these two policyholders' loss experiences identically in the modelling process. This motivates the need of the concept of *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, based on 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 known as the rate or intensity per unit period of the Poisson process and t represents the length of time or *exposure*, a known constant value. 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. For workers compensation, the rate may be the probability of injury in the course of employment per dollar and the exposure is the payroll amount in dollar.
4. 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.
5. 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.
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.

Actuaries may be able to use different exposure bases for a given insurable loss. For example, in auto insurance, both the number of kilometres driven and the number of months coved by insurance can be used as exposure bases. Here the former is more accurate and useful in modelling the losses from car accidents, but more difficult to measure and manage for insurers. Thus, a good exposure base may not be the theoretically best one due to various practical constraints. As a rule, an exposure base must be easy to determine, accurately measurable, legally and socially acceptable, and free from potential manipulation by policyholders.

Incorporating exposure in Poisson regression

As exposures affect the Poisson mean, constructing Poisson regressions requires us to carefully separate the rate and exposure in the modelling process. Focusing on the insurance context, let us 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

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

which gives

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

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 the log of exposure, $\log m_i$, is commonly called the *offset*.

2.4 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 (26) by differentiating the log-likelihood (23).

3 Categorical variables and multiplicative tariff

In this section you will learn:

- The multiplicative tariff model when the rating factors are categorical.
- How to construct the Poisson regression model based on the multiplicative tariff structure.

3.1 Rating factors and tariff

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 we can imagine that, with a small number of rating factors, there will be many policyholders falling into the same risk class, charged with the same premium. For the remaining of this chapter we assume that all rating factors are categorical variables.

To illustrate how categorical variables are used in the pricing process, we consider a hypothetical auto insurance with only two rating factors:

- Type of vehicle: Type A (personally owned) and B (owned by corporations). We use index $j = 1$ and 2 to respectively represent each level of this rating factor.
- Age band of the 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 factor.

From this classification rule, we may create an organized table or list, such as the one shown in Table 2, collected from all policyholders. Clearly there are $2 \times 3 = 6$ different risk classes in total. Each row of the table shows a combination of different risk characteristics of individual policyholders. Our goal is to compute six different premiums for each of these combinations. Once the premium for each row has been determined using the given exposure and claim counts, the insurer can replace the last two columns in Table 2 with a single column containing the computed premiums. This new table then can serve as a manual to determine the premium for a new policyholder given the rating factors during the underwriting process. In non-life insurance, a table (or a set of tables) or list that contains each set of rating factors and the associated premium is referred to as a *tariff*. Each unique combination of the rating factors in a tariff is called a *tariff cell*; thus, in Table 2 the number of tariff cells is six, same as the number of risk classes.

Table 2: Loss record of the illustrative auto insurer

Rating factors		Exposure	Claim count
Type (j)	Age (k)	in year	observed
j =1	k =1	89.1	9
1	2	208.5	8
1	3	155.2	6
2	1	19.3	1
2	2	360.4	13
2	3	276.7	6

Let us now look at the loss information in Table 2 more closely. The exposure in each row represents the sum of the length of insurance coverages, or in-force times, in the unit of year, of all the policyholders in that tariff cell. Similarly the claim counts in each row is the number of claims at each cell. Naturally 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.

In light of the Poisson regression framework, 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

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

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. That is, we have $z_{jk} = \hat{\lambda}_{jk}$ where λ_{jk} is the Poisson rate parameter. Producing z_{ij} values however does not do much beyond comparing the average loss frequencies across risk classes. To fully exploit the dataset, we will construct a pricing model from Table 2 using the Poisson regression, for the remaining part of the chapter.

We comment that actual loss records used by insurers typically include much more risk factors, in which case the number of cells grows exponentially. The tariff would then consist of a set of tables, instead of one, separated by some of the basic rating factors, such as sex or territory.

3.2 Multiplicative tariff model

In this subsection, we introduce the multiplicative tariff model, a popular pricing structure that can be naturally used within the Poisson regression framework. The developments here is based on Table 2. Recall that the loss count of a policyholder is described by the Poisson regression model with rate λ and the exposure m, so that the expected loss count becomes $m\lambda$. As m is a known constant, we are essentially concerned with modelling λ, so that it responds to the change in the rating factors. Among other possible functional forms, we commonly choose the multiplicative[2] relation to model the Poisson rate λ_{jk} for rating factor (j, k):

$$\lambda_{jk} = f_0 \times f_{1j} \times f_{2k}, \qquad j = 1, 2;\ k = 1, 2, 3. \tag{18}$$

Here $\{f_{1j}, j = 1, 2\}$ are the parameters associated with the two levels in the first rating factor, car type, and $\{f_{2k}, k = 1, 2, 3\}$ associated with the three levels in the age band, the second rating factor. For instance, the Poisson rate for a mid-aged policyholder with a Type B vehicle is given by $\lambda_{22} = f_0 \times f_{12} \times f_{22}$. The first term f_0 is some base value to be discussed shortly. Thus these six parameters are understood as numerical representations of the levels within each rating factor, and are to be estimated from the dataset.

The multiplicative form (18) is easy to understand and use, because it clearly shows 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 a vehicle of type B would be 20% larger than type A, when the other factors are the same. In non-life insurance, the parameters f_{1j} and f_{2k} are known as *relativities* as they determine how much expected loss should change relative to the base value f_0. The idea of relativity is quite convenient in practice, as we can decide the premium for a policyholder by simply multiplying a series of corresponding relativities to the base value. Dropping an existing rating factor or adding a new one is also transparent with this multiplicative structure. In addition, the insurer may easily adjust the overall premium for all policyholders by controlling the base value f_0 without changing individual relativities. However, by adopting the multiplicative form, we implicitly assume that there is no serious interaction among the risk factors.

When the multiplicative form is used we need to address an identification issue. That is, for any $c > 0$, we can write

$$\lambda_{jk} = f_0 \times \frac{f_{1j}}{c} \times c\, f_{2k}.$$

By comparing with (18), we see that the identical rate parameter λ_{jk} can be obtained for very different individual relativities. This over-parametrization, meaning that many different sets of parameters arrive at the identical model, obviously calls for some restriction on f_{1j}

[2]Preferring the multiplicative form to others (e.g., additive one) was already hinted in (4).

and f_{2k}. The standard practice is to make one relativity in each rating factor equal to one. This can be made arbitrarily, so we will assume that $f_{11} = 1$ and $f_{21} = 1$ for our purpose. This way all other relativities are uniquely determined. The tariff cell $(j, k) = (1, 1)$ is then called the *base tariff cell*, where the rate simply becomes $\lambda_{11} = f_0$, corresponding to the base value according to (18). Thus the base value f_0 is generally interpreted as the Poisson rate of the base tariff cell.

Again, (18) is log-transformed and rewritten as

$$\log \lambda_{jk} = \log f_0 + \log f_{1j} + \log f_{2k}, \tag{19}$$

as it is easier to work with in estimating process, similar to (14). This log linear form makes the log relativities of the base level in each rating factor equal to zero, i.e., $\log f_{11} = \log f_{21} = 0$, and leads to the following alternative, more explicit expression for (19):

$$\log \lambda = \begin{cases} \log f_0 + 0 + 0 & \text{for a policy in cell } (1,1), \\ \log f_0 + 0 + \log f_{22} & \text{for a policy in cell } (1,2), \\ \log f_0 + 0 + \log f_{23} & \text{for a policy in cell } (1,3), \\ \log f_0 + \log f_{12} + 0 & \text{for a policy in cell } (2,1), \\ \log f_0 + \log f_{12} + \log f_{22} & \text{for a policy in cell } (2,2), \\ \log f_0 + \log f_{12} + \log f_{23} & \text{for a policy in cell } (2,3). \end{cases} \tag{20}$$

This clearly shows that the Poisson rate parameter λ varies across different tariff cells, with the same log linear form used in the Poisson regression framework. In fact the reader may see that (20) is an extended version of the early expression (6) with multiple risk factors and that the log relativities now play the role of β_i parameters. Therefore all the relativities can be readily estimated via fitting a Poisson regression with a suitably chosen set of indicator variables.

3.3 Poisson regression for multiplicative tariff

Indicator variables for tariff cells

We now explain how the relativities can be incorporated in the Poisson regression. As seen early in this chapter we use indicator variables to deal with categorial variables. For our illustrative auto insurer, therefore, we define an indicator variable for the first rating factor as

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

For the second rating factor, we employ two indicator variables for the age band, that is,

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

and

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

The triple (x_1, x_2, x_3) then can effectively and uniquely determine each risk class. By observing that the indicator variables associated with Type A and Age band 1 are omitted, we see that tariff cell $(j, k) = (1, 1)$ plays the role of the base cell. We emphasize that our choice of the three indicator variables above has been carefully made so that it is consistent with the choice of the base levels in the multiplicative tariff model in the previous subsection (i.e., $f_{11} = 1$ and $f_{21} = 1$).

With the proposed indicator variables we can rewrite the log rate (19) as

$$\log \lambda = \log f_0 + \log f_{12} \times x_1 + \log f_{22} \times x_2 + \log f_{23} \times x_3, \tag{21}$$

which is identical to (20) when each triple value is actually applied. For example, we can verify that the base tariff cell $(j, k) = (1, 1)$ corresponds to $(x_1, x_2, x_3) = (0, 0, 0)$, and in turn produces $\log \lambda = \log f_0$ or $\lambda = f_0$ in (21) as required.

Poisson regression for the tariff model

Under this specification, let us consider n policyholders in the portfolio with the ith policyholder's risk characteristic given by a vector of explanatory variables $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})'$, for $i = 1, \ldots, n$. We then recognize (21) as

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

where $\beta_0, \ldots, \beta_3$ can be mapped to the corresponding log relativities in (21). This is exactly the same setup as in (17) except for the exposure component. Therefore, by incorporating the exposure in each risk class, the Poisson regression model for this multiplicative tariff model finally becomes

$$\log \mu_i = \log \lambda_i + \log m_i = \log m_i + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} = \log m_i + \mathbf{x}_i' \beta,$$

for $i = 1, \ldots, n$. As a result, the relativities are given by

$$f_0 = e^{\beta_0}, \quad f_{12} = e^{\beta_1}, \quad f_{22} = e^{\beta_2} \quad \text{and} \quad f_{23} = e^{\beta_3}, \tag{22}$$

with $f_{11} = 1$ and $f_{21} = 1$ from the original construction. For the actual dataset, β_i, $i = 0, 1, 2, 3$, is replaced with the mle b_i using the method in the technical supplement at the end of this chapter (Section 5).

3.4 Numerical examples

We present two numerical examples of the Poisson regression. In the first example we construct a Poisson regression model from Table 2, which is a dataset of a hypothetical auto insurer. The second example uses an actual industry dataset with more risk factors. As our purpose is to show how the Poisson regression model can be used under a given classification rule, we are not concerned with the quality of the Poisson model fit in this chapter.

Example 1: Poisson regression for the illustrative auto insurer.
In the last few subsections we considered a dataset of a hypothetical auto insurer with two risk factors, as given in Table 2. We now apply the Poisson regression model to this dataset. As done before, we have set $(j, k) = (1, 1)$ as the base tariff cell, so that $f_{11} = f_{21} = 1$. The result of the regression gives the coefficient estimates $(b_0, b_1, b_2, b_3) = (-2.3359, -0.3004, -0.7837, -1.0655)$, which in turn produces the corresponding relativities

$$f_0 = 0.0967, \quad f_{12} = 0.7405, \quad f_{22} = 0.4567 \quad \text{and} \quad f_{23} = 0.3445.$$

from the relation given in (22). The R script and the output are given below:

```
> mydat1<- read.csv("eg1_v1a.csv")
> mydat1
  Vtype Agebnd Expsr Claims
1     1      1  89.1      9
2     1      2 208.5      8
3     1      3 155.2      6
4     2      1  19.3      1
5     2      2 360.4     13
6     2      3 276.7      6
> VtypeF <- relevel(factor(Vtype), ref="1") # treat Vtype as factors with 1 as base.
> AgebndF <- relevel(factor(Agebnd), ref="1") # treat Age band as factors.
> Pois_reg1 = glm(Claims ~ VtypeF + AgebndF,
                  data = mydat1, family = poisson(link = log), offset = log(Expsr) )
> Pois_reg1

Coefficients:
(Intercept)      VtypeF2     AgebndF2     AgebndF3
    -2.3359      -0.3004      -0.7837      -1.0655

Degrees of Freedom: 5 Total (i.e. Null);  2 Residual
Null Deviance:     8.774
Residual Deviance: 0.6514  AIC: 30.37
```

Example 2: Poisson regression for Singapore insurance claims data
This actual data is a subset of the data used by Frees and Valdez (2008). The data is from the General Insurance Association of Singapore, an organisation consisting of non-life insurers in Singapore. The data contains the number of car accidents for $n = 7,483$ auto insurance policies with several categorical explanatory variables and the exposure for each policy. The explanatory variables include four risk factors: the type of the vehicle insured (either automobile (A) or other (O), denoted by `Vtype`), the age of the vehicle in years (`Vage`), gender of the policyholder (`Sex`) and the age of the policyholder (in years, grouped into seven categories, denoted `Age`). Based on the data description, there are several things to remember before constructing a model (May need the table from the Jed's pdf file). First,

there are 3,842 policies with vehicle type A (automobile) and 3,641 policies with other vehicle types. However, age and sex information is available for the policies of vehicle type A only; the drivers of all other types of vehicles are recorded to be aged 21 or less with sex unspecified, except for one policy, indicating that no driver information has been collected for non-automobile vehicles. Second, type A vehicles are all classified as private vehicles and all the other types are not.

When we include these risk factors, we assume that all unspecified sex to be male. As the age information is only applicable to type A vehicles, we set the model accordingly. That is, we apply the age variable only to vehicles of type A. Also we used five vehicle age bands, simplifying the original seven bands, by combining vehicle ages 0,1 and 2; the combined band is marked as level 2[3]. Thus our Poisson model has the following explicit form:

$$\log \mu_i = \mathbf{x}_i'\beta + \log m_i = \beta_0 + \beta_1 I(Sex_i = M) + \sum_{t=2}^{6} \beta_t \, I(Vage_i = t+1)$$
$$+ \sum_{t=7}^{13} \beta_t \, I(Vtype_i = A) \times I(Age_i = t-7) + \log m_i.$$

The fitting result is given in Table 3, for which we have several comments.

- The claim frequency is higher for male by 17.3%, when other rating factors are held fixed. However, this may have been affected by the fact that all unspecified sex has been assigned to male.

- Regarding the vehicle age, the claim frequency gradually decreases as the vehicle gets old, when other rating factors are held fixed. The level starts from 2 for this variable but, again, the numbering is nominal and does not affect the numerical result.

- The policyholder age variable only applies to type A (automobile) vehicle, and there is no policy in the first age band. We may speculate that younger drivers less than age 21 drive their parents' cars rather than having their own because of high insurance premiums or related regulations. The missing relativity may be estimated by some interpolation or the professional judgement of the actuary. The claim frequency is the lowest for age band 3 and 4, but gets substantially higher for older age bands, a reasonable pattern seen in many auto insurance loss datasets.

- We also note that there is no base level in the policyholder age variable, in the sense that no relativity is equal to 1. This is because the variable is only applicable to vehicle type A. This does not cause a problem numerically, but one may set the base relativity as follows if necessary for other purposes. Since there is no policy in age band 0, we consider band 1 as the base case. Specifically, we treat its relativity as a product of 0.918 and 1, where the former is the common relativity (that is, the common premium reduction) applied to all policies with vehicle type A and the latter is the base value for age band 1. Then the relativity of age band 2 can be seen as $0.917 = 0.918 \times 0.999$,

[3] corresponding to `VAgecat1` in the data file

where 0.999 is understood as the relativity for age band 2. The remaining age bands can be treated similarly.

Table 3: Singapore insurance claims data

Rating factor	Level	Relativity in the tariff	Note
Base value		0.167	f_0
Sex	1 (F)	1.000	Base level
	2 (M)	1.173	
Vehicle age	2 (0-2 yrs)	1.000	Base level
	3 (3-5 yrs)	0.843	
	4 (6-10 yrs)	0.553	
	5 (11-15 yrs)	0.269	
	6 (16+ yrs)	0.189	
Policyholder age	0 (0-21)	N/A	No policy
(Only applicable to	1 (22-25)	0.918	
vehicle type A)	2 (26-35)	0.917	
	3 (36-45)	0.758	
	4 (46-55)	0.632	
	5 (56-65)	1.102	
	6 (65+)	1.179	

Let us try several examples based on Table 3. Suppose a male policyholder aged 40 who owns a 7-year-old vehicle of type A. The expected claim frequency for this policyholder is then given by

$$\lambda = 0.167 \times 1.173 \times 0.553 \times 0.758 = 0.082.$$

As another example consider a female policyholder aged 60 who owns a 3-year-old vehicle of type O. The expected claim frequency for this policyholder is

$$\lambda = 0.167 \times 1 \times 0.843 = 0.141.$$

Note that for this policy the age band variable is not used as the vehicle type is not A. The R script is given below.

```
mydat <- read.csv("SingaporeAuto.csv",  quote = "", header = TRUE)
attach(mydat)

# create vehicle type as factor
TypeA = 1 * (VehicleType == "A")
table(VehicleType)
VtypeF <- as.character(VehicleType)
VtypeF[VtypeF != "A"] <- "O"
VtypeF = relevel(factor(VtypeF), ref="A")
```

```
# create gender as factor
Female = 1 * (SexInsured == "F" )
Sex = as.character(SexInsured)
Sex[Sex != "F"] <- "M"
SexF = relevel(factor(Sex), ref = "F")

# create driver age as factor
AgeCat = pmax(AgeCat - 1, 0)
AgeCatF = relevel(factor(AgeCat), ref = "0")
table(AgeCatF) # No policy in the first age band

# create vehicle age as factor
VAgeCatF = relevel( factor(VAgeCat), ref = "0" )
VAgecat1 = factor(VAgecat1, labels =
              c("Vage0-2", "Vage3-5", "Vage6-10", "Vage11-15", "Vage15+") )
VAgecat1F = relevel( factor(VAgecat1), ref = "Vage0-2" )

# Poisson reg model
Pois_reg2 = glm(Clm_Count ~ SexF + TypeA:AgeCatF + VAgecat1F,
                  offset = LNWEIGHT, poisson(link = log) )
summary(Pois_reg2)

# compute relativities
exp(Pois_reg2$coefficients)

detach(mydat)
```

4 Further Reading and References

The Poisson regression is a special member of a more general regression model class known as the generalized linear model (glm). The glm develops a unified regression framework for datasets when the response valuables are continuous, binary or discrete. The classical linear regression model with normal error is also a member of the glm. There are many standard statistical texts dealing with the glm, including McCullagh and Nelder (1989). More accessible texts are Dobson and Barnett (2008), Agresti (1996) and Faraway (2016). For actuarial and insurance applications of the glm see Frees (2010), De Jong and Heller (2008). Also, Ohlsson and Björn (2010) discusses the glm in non-life insurance pricing context with tariff analyses.

5 Technical supplements – Estimating Poisson regression model

Maximum likelihood estimation for individual data

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. Again, we assume that there are n policyholders and the ith policyholder is characterized by $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik})'$ with the observed loss count y_i. Then, from (16) and (17), the log-likelihood function of vector $\beta = (\beta_0, \ldots, \beta_k)$ is given by

$$\begin{aligned} \log L(\beta) = l(\beta) &= \sum_{i=1}^{n} \left(-\mu_i + y_i \log \mu_i - \log y_i!\right) \\ &= \sum_{i=1}^{n} \left(-m_i \exp(\mathbf{x}_i'\beta) + y_i \left(\log m_i + \mathbf{x}_i'\beta\right) - \log y_i!\right) \end{aligned} \tag{23}$$

To obtain the mle of $\beta = (\beta_0, \ldots, \beta_k)'$, we differentiate[4] $l(\beta)$ with respect to vector β and set it to zero:

$$\frac{\partial}{\partial \beta} l(\beta)\bigg|_{\beta=\mathbf{b}} = \sum_{i=1}^{n} \left(y_i - m_i \exp(\mathbf{x}_i'\mathbf{b})\right) \mathbf{x}_i = \mathbf{0}. \tag{24}$$

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 (24) 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 (24) as

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

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

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

which implies that we must have

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

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 sample mean is preserved under the fitted Poisson regression model.

[4] We use matrix derivative here.

Maximum likelihood estimation for grouped data

Sometimes the data is not available at the individual policy level. For example, Table 2 provides collective loss information for each risk class after grouping individual policies. When this is the case, y_i and m_i, the quantities needed for the mle calculation in (24), are unavailable for each i. However this does not pose a problem as long as we have the total loss counts and total exposure for each risk class.

To elaborate, let us assume that there are K different risk classes, and further that, in the kth risk class, we have n_k policies with the total exposure $m_{(k)}$ and the average loss count $\bar{y}_{(k)}$, for $k = 1, \ldots, K$; the total loss count for the kth risk class is then $n_k\, \bar{y}_{(k)}$. We denote the set of indices of the policies belonging to the kth class by C_k. As all policies in a given risk class share the same risk characteristics, we may denote $\mathbf{x}_i = \mathbf{x}_{(k)}$ for all $i \in C_k$. With this notation, we can rewrite (24) as

$$\begin{aligned}
\sum_{i=1}^{n} \left(y_i - m_i \exp(\mathbf{x}_i'\mathbf{b})\right) \mathbf{x}_i &= \sum_{k=1}^{K} \Big\{ \sum_{i \in C_k} \left(y_i - m_i \exp(\mathbf{x}_i'\mathbf{b})\right) \mathbf{x}_i \Big\} \\
&= \sum_{k=1}^{K} \Big\{ \sum_{i \in C_k} \left(y_i - m_i \exp(\mathbf{x}_{(k)}'\mathbf{b})\right) \mathbf{x}_{(k)} \Big\} \\
&= \sum_{k=1}^{K} \Big\{ \Big(\sum_{i \in C_k} y_i - \sum_{i \in C_k} m_i \exp(\mathbf{x}_{(k)}'\mathbf{b}) \Big) \mathbf{x}_{(k)} \Big\} \\
&= \sum_{k=1}^{K} \Big(n_k\, \bar{y}_{(k)} - m_{(k)} \exp(\mathbf{x}_{(k)}'\mathbf{b}) \Big) \mathbf{x}_{(k)} = 0. \qquad (25)
\end{aligned}$$

Since $n_k\, \bar{y}_{(k)}$ in (25) represents the total loss count for the kth risk class and $m_{(k)}$ is its total exposure, we see that for the Poisson regression the mle $\mathbf{b}$ is the same whether if we use the individual data or the grouped data.

Information matrix

Taking second derivatives to (23) gives the information matrix of the mle estimators,

$$\mathbf{I}(\beta) = -\mathrm{E}\left(\frac{\partial^2}{\partial \beta \partial \beta'} l(\beta) \right) = \sum_{i=1}^{n} m_i \exp(\mathbf{x}_i'\beta) \mathbf{x}_i \mathbf{x}_i' = \sum_{i=1}^{n} \mu_i \mathbf{x}_i \mathbf{x}_i'. \qquad (26)$$

For actual datasets, μ_i in (26) 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.

For grouped datasets, we have

$$\mathbf{I}(\beta) = \sum_{k=1}^{K} \Big\{ \sum_{i \in C_k} m_i \exp(\mathbf{x}_i'\beta) \mathbf{x}_i \mathbf{x}_i' \Big\} = \sum_{k=1}^{K} m_{(k)} \exp(\mathbf{x}_{(k)}'\beta) \mathbf{x}_{(k)} \mathbf{x}_{(k)}'.$$

References

Agresti, A. (1996). *An introduction to categorical data analysis*, Vol. 135. Wiley New York.

De Jong, P. and G. Z. Heller (2008). *Generalized linear models for insurance data*, Vol. 10. Cambridge University Press Cambridge.

Dobson, A. J. and A. Barnett (2008). *An introduction to generalized linear models*. CRC press.

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*, Vol. 124. CRC press.

Frees, E. W. (2010). *Regression modeling with actuarial and financial applications*, New York. Cambridge University Press.

Frees, E. W. and E. A. Valdez (2008). "Hierarchical insurance claims modeling," *Journal of the American Statistical Association*, Vol. 103, pp. 1457–1469.

McCullagh, P. and John A. Nelder (1989). *Generalized Linear Models*. Chapman & Hall, 2nd edition.

Ohlsson, E. and J. Björn (2010). *Non-life insurance pricing with generalized linear models*, Vol. 21. Springer.