Ch 8. Risk classification

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

February 27, 2018

Chapter Preview. To be done.

1 Introduction

Through insurance contracts, the policyholders effectively transfer these 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 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.

However, this individual pricing would be impractical in reality because identifying, measuring and incorporating all relevant risk characteristics are quite cumbersome, difficult and expensive for insurers; some relevant risk characteristics, e.g., driving skills, are not even directly observable. Instead, insurers maintain some classification system which assigns each policyholder to one of the risk classes based on a relatively small number of risk characteristics that are deemed more relevant. These characteristics used in the classification system are called the rating factors; these factors are *a priori* variables in the sense that they are known before the contract begins. Therefore, a risk classification system can be thought of as a compromise between the two extreme premiums: the true individual premium and the collective premium where all policyholders pay the same premium regardless of their risk characteristics. All policyholders in the same risk class therefore pay the same premium as they share identical risk factors.

An important task in this 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 a risk classification scheme.

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 the important quantities, which could provide a better intuition. We start our discussion with the standard linear regression model.

2.1 Linear regression

Suppose that we have a single a priori rating factor denoted x for each policyholder along with the past loss amount, which is also a single number y. Then for a pool of n independent policyholders the insurer's dataset consists of (x_i, y_i), $i = 1, \ldots, n$, where the subscript indicates the ith policyholder. The simplest linear regression model then postulates a linear relation between x and y, so that

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

Here β_0 and β_1 are the common intercept and slope coefficients which apply to all policyholders. The last term ϵ_i, known as the random error term, represents the random uncertainty that cannot be captured by the pure deterministic equation $y = \beta_0 + \beta_1 x$. It is commonly assumed that $\epsilon_i \sim_{iid} N(0, \sigma^2)$, so that we have, from (1),

$$\mathrm{E}(y_i|x_i) = \beta_0 + \beta_1 x_i. \tag{2}$$

A generic form of this model without the subscripts, called the regression or mean response function, is written as

$$\mathrm{E}(y|x) = \beta_0 + \beta_1 x \tag{3}$$

Thus the *expected* loss $\mathrm{E}(y|x)$ is a linear function of the rating factor x. In the statistical literature, the input x is often referred to as the predictor or explanatory variable, and the output y as the response variable. This functional form indicates that when the rating factor x increases by 1, the loss amount increases by β_1 on average. This formulation however is merely a theoretical one because the true values of β_0 and β_0 are unknown. To obtain a workable model, we estimate these parameters using, e.g., maximum likelihood estimation (mle)[1] based on the normal distributional assumption for ϵ_i. By denoting these estimates by $\hat{\beta}_0 = b_0$ and $\hat{\beta}_1 = b_1$, we obtain the estimated regression function

$$\hat{y}_i = b_0 + b_1 x_i, \tag{4}$$

[1]For the linear regression, the mle and ols yield the same estimators.

paralleling Equation (2). Thus, if the postulated linear model in (1) is accepted, we can estimate the expected loss for the ith policyholders simply as $\hat{y}_i = b_0 + b_1 x_i$, and policyholders with a different rating factor produces a different premium.

When there are more than one rating factors, as is the case in practice, we extend this simple linear form (1) into a multiple linear regression. When there are $k \geq 1$ different rating factors for each policyholder, the record of the ith policyholder would consist of $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$, $i = 1, \ldots, n$, and the multiple linear regression postulates

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \ldots, n. \tag{5}$$

By using matrix notation we can compactly rewrite this model as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{6}$$

where each vector and matrix are given by

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_2 & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_{n,k} \end{pmatrix}$$

The corresponding regression function is then given by

$$\mathrm{E}(y|\mathbf{x}) = \mathbf{x}'\beta \tag{7}$$

where $\mathbf{x} = (1, x_1, \ldots, x_k)'$ is the vector of some hypothesized rating factors. As before, the parameter vector β must be estimated from the dataset, and one can obtain the estimated regression function

$$\hat{y} = b_0 + b_1 x_1 + \ldots + b_k x_k = \mathbf{x}'\mathbf{b}, \tag{8}$$

where $\mathbf{b} = (b_0, \ldots, b_k)' = (\hat{\beta}_0, \ldots, \hat{\beta}_k)'$.

The linear regression models are important statistical tools in data analyses and encompass a wide variety of extensions but we do not delve into this too much as our goal here in this chapter is to introduce an alternative regression model that can handle count response variables.

2.2 Need of Possion regression model

Now suppose that we have a dataset where the response variable is the loss count instead of the loss amount. For example, y can be the number of accidents of a policyholder during a year. In this case the loss can take only non-negative integer values, i.e., $y_i = 0, 1, 2, \ldots$ for each policyholder. Unfortunately, the multiple linear regression above, despite being popular and useful in many applications, cannot handle this type of dataset because of the following reasons. First, there is no guarantee that the output $\hat{y}$ in (4) is an integer value. It is easy to imagine that it can be a fractional number by adjusting some rating factors. Second, more importantly, there is possibility that $\hat{y}$ becomes negative for a peculiar set of rating

factors. These unintended and undesirable consequences are actually due to the fact that the output of multiple linear regression models spans the whole real line, that is, $\hat{y}$ can take any real value in $(-\infty, \infty)$ by construction. More fundamentally, this stems from the very assumption of the normal error terms in the regression model itself. This can be readily verified by observing that, when $\epsilon_i \sim_{iid} N(0, \sigma^2)$, the response variable is also normally distributed according to (1):

$$y_i \sim_{iid} N(\beta_0 + \beta_1 x_i, \sigma^2), \tag{9}$$

which confirms that the loss amount y is allowed to take any real value.

One may impose restrictions on the linear regression so that it can take only non-negative integers as the response variable. Instead of seeking such ad hoc solutions however we consider the Poisson regression, a different type of regression model where the underlying random component is consistent with the count variable. We have encountered the Poisson distribution and studied its distributional properties in early chapters. The pmf of the Poisson with mean μ is

$$\Pr(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, \qquad y = 0, 1, 2, \ldots \tag{10}$$

and $\mathrm{E}(Y) = \mathrm{Var}(Y) = \mu$. Now suppose Y stands for the random loss count, and it depends on k rating factors $\mathbf{x} = (1, x_1, \ldots, x_k)$, as before. Our goal is then to formulate a regression model which uses the Poisson distribution as the response variable and, at the same time, responds to risk factors of individual policyholders.

2.3 Formulaing Poisson regression

As before we assume that n independent policyholders with rating factors $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik})$ have recorded the loss count $y_i = 0, 1, 2, \ldots$, for $i = 1, \ldots, n$. We also assume, for now, that all policyholders have the same one unit period for loss observation, or equal exposure[2] of 1, to make things simpler. We start with looking at the linear regression model from a different angle. The linear regression formulated in (1) and (3), when combined, suggests an alternative description of the model

$$y_i = \mathrm{E}(y_i|\mathbf{x}_i) + \epsilon_i \tag{11}$$

with $\mathrm{E}(y_i|\mathbf{x}_i) = \mathbf{x}_i'\beta$, where $\beta = (\beta_0, \beta_1, \ldots, \beta_k)'$. This expression tells that the response variable is the mean response function with the error term added. Along similar lines, we can 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 (10):

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

It is noted here that in principle the conditional mean $\mathrm{E}(y_i|\mathbf{x}_i)$ can take different forms depending on how we specify the relationship between $\mathbf{x}$ and y, though it was chosen as a

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

linear function in the linear regression. The standard choice for the Poisson regression is to adopt the exponential function so that

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

Note that this specification resolves the limitations of the linear regression model previously mentioned. The exponential function ensures that $\mu_i > 0$ for any set of rating factors $\mathbf{x}_i$ and each y_i appropriately forms count data generated from 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.

2.4 Incorporating exposure

The assumption of the same one unit of exposure across all policyholders is not a realistic one. In auto insurance, 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 (10) is parametrised via its mean. To understand the exposure, we alternatively parametrize the Poisson pmf in terms of the *rate* parameter λ, inspired from 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. As the exposure t gets larger, the Poisson distribution produces a larger expected loss count as it should. Clearly, (15) reduces to (10) 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.

Incorporating the exposure in the Poisson regression thus requires us to carefully separate the rate and exposure in the modelling process. Focusing on our 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*.

2.5 Estimating Poisson regression model

2.6 Exercises

1. Show that the mle and ols are the same for the simple linear regression.

2. In the Poisson regression formulation, 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.

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 1, 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.

Table 1: 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

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.

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

$$z_{jk} = \frac{y_{jk}}{m_{jk}}, \qquad j = 1, 2;\ k = 1, 2, 3. \tag{18}$$

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 1, 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,

$$\lambda_{jk} = f_0 \times f_{1j} \times f_{2k}, \qquad j = 1, 2;\ k = 1, 2, 3. \tag{19}$$

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 (19) 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, (19) has an identification issue because, for any $c > 0$, we have

$$\lambda_{ij} = f_0 \times \frac{f_{1j}}{c} \times c\, f_{2k}, \tag{20}$$

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 (19). 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 (19) is log-transformed and rewritten as

$$\log \lambda_{jk} = \log f_0 + \log f_{1j} + \log f_{2k}, \tag{21}$$

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 (21) 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

$$x_1 = \begin{cases} 1 & \text{for vehicle type B,} \\ 0 & \text{for vehicle type A.} \end{cases} \tag{22}$$

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[3], that is,

$$x_2 = \begin{cases} 1 & \text{for age band 2,} \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

and

$$x_3 = \begin{cases} 1 & \text{for age band 3,} \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

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 (21) in terms of indicator variables as

$$\log \lambda_{jk} = \log f_0 + \log f_{12} \times x_1 + \log f_{22} \times x_2 + \log f_{23} \times x_3. \tag{25}$$

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 (25), $\log \lambda_{22} = \log f_0 + \log f_{22}$, 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 (25), 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 (25) as $\log f_{11} = \log f_{21} = 0$ by construction. If we reverse the binary switch for x_i in (22) – (24) or select different coefficients in (25), 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 1. Then for the ith risk class with triple (x_{i1}, x_{i2}, x_{i3}), we now recognize (25) as

$$\log \lambda_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} = \mathbf{x}_i' \beta, \qquad i = 1, \ldots, 6, \tag{26}$$

where $\beta_0, \ldots, \beta_3$ can be found in the corresponding position in (25). 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 2 we present the mapping between β_i coefficients and the price relatives, keeping the same list form as Table 1. All columns except for the first is indexed with i, giving us a complete specification for our Poisson regression.

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

Table 2: 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

$$\log \mu_i = \mathbf{x}_i \beta \tag{27}$$

or

$$\mu_i = e^{\mathbf{x}_i \beta} \tag{28}$$

- $$y|\mathbf{x} = f(\mathbf{x}, \beta) + \varepsilon \tag{29}$$

  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

$$E(y|\mathbf{x}) = f(\mathbf{x}, \beta) := \mu \tag{30}$$

- 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

$$\mu_i = E(y_i|\mathbf{x}_i) = f(\mathbf{x}_i, \beta) \tag{31}$$
$$\tag{32}$$

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

5 Application

6