UNIVERSITY OF WATERLOO

ACTSC 632 Spring 2023

Data Science with Actuarial Applications

Week 6

Xintong Li

Department of Statistics and Actuarial Science

Last Week

- Y is a key ratio: claim frequency or claim severity
- X is a vector of rating factors, modeled as categorical variables (can still use dummy binary variables to implement this in R)
- terminology: duration, claim frequency, claim severity, rating cell
- 3 key assumptions
- $\mathsf{E}(Y) = \mu$, $\mathsf{Var}(Y) = \sigma^2/w$
- Moped example (in R)

Today

- Introduce the idea of multiplicative models
- Basic model for claim frequency
- Basic model for claim severity
- How will we use GLMs

Note on Multiplicative Models

2.4.1 Basic model for claim frequency

Reproductive property of the relative Poisson distribution

2.4.2 Basic model for claim severity

For both the frequency and the severity, the impact of the rating factors will be incorporated via μ_i, using the theory of GLMs.

2.5 The basics of pricing with GLMs

- **Goal:** determine how key ratio Y varies with rating factors
- Q: Why not use multiple linear regression?
  - (A) Assumption of normally distributed error may not be reasonable, e.g., for nb of claims (which is discrete) or claim size (which is skewed and > 0)
  - (B) modeling Y as a linear (additive) fct of the rating factors clashes with our intuition to use a multiplicative model
  - (C) error term forced to have constant variance
- **Advantages of GLM for pricing**
  1. has theory that can be used to estimate std error, build CIs, do model selection
  2. used in many areas, so can benefit for developments elsewhere
  3. std software is available for fitting

Quick review of GLMs

What is next

- Next we discuss how **Exponential Dispersion Models (EDM)** can be used to address (A) (Assumption of normally distributed error may not be reasonable) in the context of non-life insurance pircing.
- Then we'll review how the flexibility in choosing the **link function** can be used to address (B) (modeling Y as a linear (additive) fct of the rating factors) and allow us to use a multiplicative model

2.5.1 Exponential Dispersion Models

Recap of Week 6 – Lecture 1

- **Basic claim frequency model**: use

$$\mathbb{P}\left(Y_i = y_i\right) = \frac{\left(w_i\mu_i\right)^{w_i y_i} e^{-w_i\mu_i}}{k!}, \qquad y_i \in \left\{0, \frac{1}{w_i}, \frac{2}{w_i}, \ldots\right\}$$

- **Basic claim severity model**: use

$$f_{Y_i}\left(y\right) = \frac{\left(\frac{\mu_i}{\xi_i} w_i\right)^{\frac{w_i\left(\mu_i\right)^2}{\xi_i}} y^{\frac{w_i\left(\mu_i\right)^2}{\xi_i}-1} e^{-\frac{w_i\mu_i}{\xi}y}}{\Gamma\left(\frac{w_i\left(\mu_i\right)^2}{\xi_i}\right)}, \qquad y_i > 0$$

Recap of Week 6 – Lecture 1

- Y_i is EDM if pdf/pmf given by

$$f_{Y_i}(y_i, \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\frac{\phi}{w_i}} + c(y_i, \phi, w_i)\right\},$$

- Forgot to provide notation $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ for the linear predictor

Additional properties of EDM

Checking that relative Poisson distribution is EDM

Checking that Gamma model is an EDM

Reproductive property of EDM family

Result: If Y_1 and Y_2 are two independent rv's from the same EDM family (i.e., same $b(\cdot)$, same μ and same ϕ) but with possibly different weights w_1 and w_2 then $Y = \frac{w_1 Y_1 + w_2 Y_2}{w_1 + w_2}$ is in same EDM family but with weight $w_1 + w_2$.

2.5.2 Link Function

- Previous topic helped us identify a rich class of models for Y that are useful in the context of GLMs
- The **link function** provides flexibility to model how the response (i.e., the key ratio Y_i) relates to the rating factors.
- This is done by imposing a **relationship** between the mean response $\mathbb{E}\left[Y_i\right] = \mu_i$ and the set of rating factors.

Example with one rating factor

- Consider a tariff model with only one rating factor taking possible values $\{a, b, c\}$.
- One category will be the base category for the rating factor (say, category a) and **2 binary variables** will be created to indicate whether the category is b or not, and whether the category is c or not. Hence we get the linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2},$$

where $x_{i1} = 1$ if tariff cell i has category b for rating factor (or otherwise 0) and $x_{i2} = 1$ if tariff cell i has category c for its rating factor (or otherwise 0). Hence

Tariff cell i	Rating factor	η_i
1	(a)	β_0
2	(b)	$\beta_0 + \beta_1$
3	(c)	$\beta_0 + \beta_2$

Example with two rating factors

Consider now a tariff model with two rating factors:

- Rating factor 1 takes on two possible values $\{a, b\}$;
- Rating factor 2 takes on three possible values $\{c, d, e\}$.
- Create 1 binary variable for Rating factor 1 (assuming base category is a)
- Create 2 binary variables for Rating factor 2 (assuming base category is c).

Hence, we define

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3},$$

where

- $x_{i1} = 1$ if tariff cell i has category b for its first rating factor (or otherwise 0);
- $x_{i2} = 1$ if tariff cell i has category d for its second rating factor (or otherwise 0)
- $x_{i3} = 1$ if tariff cell i has category e for its second rating factor (or otherwise 0)

Tariff cell i	Rating factors	η_i
1	(a, c)	β_0
2	(a, d)	$\beta_0 + \beta_2$
3	(a, e)	$\beta_0 + \beta_3$
4	(b, c)	$\beta_0 + \beta_1$
5	(b, d)	$\beta_0 + \beta_1 + \beta_2$
6	(b, e)	$\beta_0 + \beta_1 + \beta_3$

Vectors of binary variables corresponding to 6 rating cells:

$$\vec{x}_1 = (1, 0, 0, 0) \qquad \vec{x}_4 = (1, 1, 0, 0)$$
$$\vec{x}_2 = (1, 0, 1, 0) \qquad \vec{x}_5 = (1, 1, 1, 0)$$
$$\vec{x}_3 = (1, 0, 0, 1) \qquad \vec{x}_6 = (1, 1, 0, 1)$$

$X = [x_{ij}]_{i=1,j=1}^{4\ \ 6}$ is the design matrix

- More generally, for a rating factor with say p categories, we shall create $(p-1)$ binary variables, each of which takes the value 1 if tariff cell i is in a given category and 0, otherwise.
- One of the categories is assumed to be the base category which is why we create only $(p-1)$ binary variables.

2.5.2 Link Function