SF2930 GLM Lecture 1

February 18, 2018

1 Introduction

Without customers we would not have any business. We always have our customers top of mind and want to make them feel safe and off course stay with us for as long as possible. So how could we find out which customers will stay with us and which will leave us? This will be the main topic of this lecture and introduce generalized linear models (GLMs) to answer this question together with the concept of the exponential family.

2 From Additive Linear Model to GLM Form

Using a linear additive model we would obtain the mean renewal rate, μ_{ij}, for our customer's policies (the insurance contracts) by

$$\mu_{ij} = \gamma_0 + \gamma_{1i} + \gamma_{2j}, \tag{1}$$

where γ_0 is a *base level*, i refers to the customer's group w.r.t. the sales channel and j refers to the customer's group w.r.t. price change from last year. γ_{1i} is thus the parameter for the ith group in *Variable 1*. There are 3 different sales channels in our data set, Call center, Face to face and Broker. Furthermore, we have divided the price change variable into 2 groups smaller and larger than 10%. The response variable for each individual insurance policy is then

$$y_i = \begin{cases} 1, & \text{if the insurance policy was renewed} \\ 0, & \text{if the insurance policy was not renewed.} \end{cases}$$

We have aggregated the data in Table 1, in which each unique combination of variable groups i, j correspond to one row, often referred to as a *cell*. For each cell there is a corresponding equation on the form of Eq. (1),

$$\begin{aligned}
\mu_{11} &= \gamma_0 + \gamma_{11} + \gamma_{21}, \\
\mu_{12} &= \gamma_0 + \gamma_{11} + \gamma_{22}, \\
\mu_{21} &= \gamma_0 + \gamma_{12} + \gamma_{21}, \\
\mu_{22} &= \gamma_0 + \gamma_{12} + \gamma_{22}, \\
\mu_{31} &= \gamma_0 + \gamma_{13} + \gamma_{21}, \\
\mu_{31} &= \gamma_0 + \gamma_{13} + \gamma_{22}.
\end{aligned}$$

Table 1: Aggregated historic insurance renewal data with three groups for Variable 1, Sales Channel, and two groups for Variable 2, Price change.

Cell	Sales channel	Price change	Number of customers	Number of renewed	Renewal rate
	Variable 1	*Variable 2*	w	X	$Y = X/w$
1	Call center (1)	$< 10\%$ (1)	12033	11260	93.6%
2	Call center (1)	$\geq 10\%$ (2)	959	763	79.6%
3	Face to face (2)	$< 10\%$ (1)	2056	1914	93.1%
4	Face to face (2)	$\geq 10\%$ (2)	108	91	84.3%
5	Broker (3)	$< 10\%$ (1)	3178	2901	91.3%
6	Broker (3)	$\geq 10\%$ (2)	231	171	74.0%

This model is over parametrized, hence, it has more parameters, γ, than needed, which gives us the freedom to define a *base cell* in which only the base level is non-zero. Choosing $(1, 1)$ as our base cell we set $\gamma_{11} = \gamma_{21} = 0$. Renaming the parameters according to

$$\begin{cases} \beta_0 & \doteq \gamma_0 \\ \beta_1 & \doteq \gamma_{12} \\ \beta_2 & \doteq \gamma_{23} \\ \beta_3 & \doteq \gamma_{22} \end{cases}$$

with which we get

$$\begin{array}{llllll} \mu_{11} & = \beta_0 & & & \\ \mu_{12} & = \beta_0 & & & +\beta_3 \\ \mu_{21} & = \beta_0 & +\beta_1 & & \\ \mu_{22} & = \beta_0 & +\beta_1 & & +\beta_3 \\ \mu_{31} & = \beta_0 & & +\beta_2 & \\ \mu_{32} & = \beta_0 & & +\beta_2 & +\beta_3, \end{array}$$

where we see that β_1 describes the difference between call center and face to face, β_2 the difference between call center and broker and β_3 between less than 10% and more than 10% price change. Renaming the mean renewal rate for cell i to μ_i and introducing zeros according to

$$\begin{array}{lllll} \mu_0 & = 1 \cdot \beta_0 & +0 \cdot \beta_1 & +0 \cdot \beta_2 & +0 \cdot \beta_3 \\ \mu_1 & = 1 \cdot \beta_0 & +0 \cdot \beta_1 & +0 \cdot \beta_2 & +1 \cdot \beta_3 \\ \mu_2 & = 1 \cdot \beta_0 & +1 \cdot \beta_1 & +0 \cdot \beta_2 & +0 \cdot \beta_3 \\ \mu_3 & = 1 \cdot \beta_0 & +1 \cdot \beta_1 & +0 \cdot \beta_2 & +1 \cdot \beta_3 \\ \mu_4 & = 1 \cdot \beta_0 & +0 \cdot \beta_1 & +1 \cdot \beta_2 & +0 \cdot \beta_3 \\ \mu_5 & = 1 \cdot \beta_0 & +0 \cdot \beta_1 & +1 \cdot \beta_2 & +1 \cdot \beta_3 \end{array}$$

we can express the system of equations in a more compact way

$$\mu_i = \sum_{j=0} x_{ij} \beta_j, \tag{2}$$

Table 2: Comparison between ordinary linear regression models and GLM

Model	Randomness	Structure
Regression model	$Y_i \sim N(\mu_i, \sigma_i)$	$\mu_i = \sum_{j=1} x_{ij}\beta_j$
GLM	$Y_i \sim P(\mu_i, \sigma_i)$	$g(\mu_i) = \sum_{j=1} x_{ij}\beta_j$

where $i = 0, 1, ..., 5$ and we have introduced the dummy variables

$$x_{ij} = \begin{cases} 1, & \text{if } \beta_j \text{ is included in } \mu_i \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

.

We have now transformed Eq. (1) to the most basic GLM form in Eq. (2), through which we also have gained fundamental knowledge on how the parameters β_i are linked to the different cells of the model and, thus, the core of GLMs.

3 GLM and Logistic Regression

Having found the correct form we still have some issues to deal with since what we are predicting can only take on values between 0 and 1. Weather or not a customer will renew its policy can be seen as the outcome of a Bernoulli trial according to

$$\begin{aligned} \Pr(y_i = 1) &= \pi_i, \\ \Pr(y_i = 0) &= 1 - \pi_i, \end{aligned} \tag{4}$$

where we have changed to the common notation of π_i instead of μ_i for this particular response variable. The expected value is then found to be

$$E[y_i] = 1 \cdot \pi_i + 0 \cdot (1 - \pi_i) = \pi_i.$$

In general, linear regression models assume that data come from a Normal distribution with the mean related to predictors. It is easy to see that this is not applicable to Eq. (4) since the error cannot be normally distributed with only the two possible outcomes. On the contrary, GLMs assume that data come from some distribution, member of the *exponential family*, with a function, g, of the mean related to predictors according to

$$g(\mu_i) = \sum_{j=0} x_{ij}\beta_j, \tag{5}$$

which is the most general form of GLMs. These main differencies are shown in Table 2.

The function, g, is called the *link function* and is the key to solving the problem with restricting π_i to the interval $[0, 1]$. This is done by using the *logit* link function according to

$$g(\pi_i) = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \sum_{j=0} x_{ij}\beta_j, \tag{6}$$

Table 3: Some of the possible link functions, the relations to the mean, μ_i, and four common distributions of the response variables that they are compatible with. $\star$ indicates that it often used as default and $\checkmark$ that it is compatible.

Link	$g(\mu_i)$	$\mu_i =$	Normal	Binomial	Poisson	Gamma
identity	μ_i	$\sum x_{ij}\beta_j$	$\star$		$\checkmark$	$\checkmark$
log	$\ln(\mu_i)$	$e^{\sum x_{ij}\beta_j}$	$\checkmark$		$\star$	$\checkmark$
inverse	$1/\mu_i$	$(\sum x_{ij}\beta_j)^{-1}$	$\checkmark$			$\star$
sqrt	$\sqrt{\mu_i}$	$(\sum x_{ij}\beta_j)^2$			$\checkmark$	
logit	$\ln(\mu_i/(1-\mu_i))$	$(1+e^{-\sum x_{ij}\beta_j})^{-1}$		$\star$		

where ln is the natural logarithm, with which

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) \in \mathbb{R},$$

and the quantity $\pi_i/\left(1-\pi_i\right) \in \mathbb{R}^+$ is called *odds*.

For GLM in general there are several possible link functions and the choice is strongly related to the distribution of the response variable. In Table 3 common link functions and compatible distributions are shown. It may seem like the there are no restrictions on the distribution of the response variable, however, as previously mentioned, a key assumption is that it is a member of the exponential family which we turn to next.

4 Exponential Family

By assuming that the variables $Y_1, ..., Y_n$ are independent, which in general is required in GLM theory, the probability distribution is given by the general form[1]

$$f_{Y_i}\left(y_i; \theta_i, \phi\right) = \exp\left\{\frac{y_i\theta_i - b\left(\theta_i\right)}{\phi/w_i} + c\left(y_i, \phi, w_i\right)\right\}, \tag{7}$$

where

- Y_i is the key ratio in cell i,
- θ_i is called the *natural location parameter* which is allowed to change with i and is related to the mean μ_i,
- ϕ is called the *dispersion parameter*, or *scale parameter*, and is the same for all cells,
- w_i is the weight of the cell, in our case the number of cusotmers,

[1] Strictly speaking we focus on Exponential Dispersion Models in this section, which is somewhat less general than the complete exponential family.

- $b(\theta_i)$ is called the *cumulant function* which has useful properties as we will see, and
- $c(y_i, \phi, w_i)$ does not depend on θ_i and is of little interest, but is required in order for the total probability to equal one.

The cumulant function, $b(\theta_i)$, is assumed to be twice continuously differentiable with invertible first derivative. For every choice of such a function we get a function we find a family of probability distributions, e.g. the ones listed in Table 3. Having set the function $b(\theta_i)$ the distribution is completely specified by the parameters θ_i and ϕ. Other technical restriction are that $\phi > 0$, $w_i \geq 0$ and that the parameter space must be open, e.g., $0 < \theta_i < 1$ which we will come back to later.

The importance of the cumulant function is seen in

$$\mu_i = E\left[y_i\right] = \frac{db(\theta_i)}{d\theta_i}, \tag{8}$$

and[2]

$$\mathrm{Var}(\mu_i) = \frac{\mathrm{Var}(y_i)}{\phi/w_i} = \frac{d\mu_i}{d\theta_i} = \frac{d^2 b(\theta_i)}{d\theta_i^2}, \tag{9}$$

for members in the exponential family. These properties stems from the *cumulant-generating function*, $\Psi(t)$, which is the logarithm of the so called *moment-generating function* which is given by

$$M(t) = E\left[e^{tY}\right],$$

where we have droped the i notation on Y_i for convenience. Let us derive the two cumulants in Eq. (8) and Eq. (9).

Using the expression for the probability distribution in Eq. (7) we find

$$\begin{aligned} E\left[e^{tY}\right] &= \int e^{tY} f_{Y_i}\left(y_i; \theta_i, \phi\right) dy \\ &= \int \exp\left(\frac{y(\theta + t\phi/w) - b(\theta)}{\phi/w} + c(y, \phi, w)\right) dy \\ &= \exp\left(\frac{b(\theta + t\phi/w) - b(\theta)}{\phi/w}\right) \\ &\qquad \times \int \exp\left(\frac{y(\theta + t\phi/w) - b(\theta + t\phi/w)}{\phi/w} + c(y, \phi, w)\right) dy, \end{aligned} \tag{10}$$

where we have multiplied with

$$1 = \exp\left(\frac{b(\theta + t\phi/w)}{\phi/w}\right) \exp\left(-\frac{b(\theta + t\phi/w)}{\phi/w}\right),$$

in the third step. Now, in the integral, we identify that it is simply the probability distribution function in Eq. (7) with $\theta \to \theta + t\phi/w$. Thus, in a neighborhood

[2] This relation for the variance holds for all but the Normal distribution.

of 0, for $|t| < \delta$ for some $\delta > 0$, $\theta + t\phi/w$ will be in the parameter space since we required the parameter space to be open. This implies that we are summing over the entire probability density function which is simply 1. Hence, we find the moment generating function

$$M(t) = \exp\left(\frac{b(\theta + t\phi/w) - b(\theta)}{\phi/w}\right),$$

and the cumulant generating function

$$\Psi(t) = \ln\left(M(t)\right) = \frac{b(\theta + t\phi/w) - b(\theta)}{\phi/w}.$$

The cumulants are then found by differentiating w.r.t. t and evaluating at 0

$$\Psi'(0) = b'(\theta) = E\left[Y\right] = \mu,$$

and

$$\Psi''(0) = b''(\theta)\phi/w = \mathrm{Var}(y),$$

and for all but the normal distribution we have that

$$\mathrm{Var}(\mu) = \frac{\mathrm{Var}(y)}{\phi/w} = \frac{d\mu_i}{d\theta_i} = \frac{d^2 b(\theta_i)}{d\theta_i^2}.$$

As an example we consider the normal distribution to make sure that it is part of the exponential family and find that $\theta_i = \mu_i, \phi = \sigma^2$ and $b(\theta_i) = \theta_i^2/2$ which yields

$$f_{Y_i}(y_i) = \exp\left\{\frac{y_i\mu_i - \mu_i^2/2}{\sigma^2/w_i} + c\left(y_i, \phi, w_i\right)\right\} \tag{11}$$

where

$$c\left(y_i, \phi, w_i\right) = -\frac{1}{2}\left(\frac{w_i y_i^2}{\sigma^2} + \log\left(2\pi\sigma^2/w_i\right)\right).$$

5 Maximum Likelihood Estimation of β_j

With an expression for f_{Y_i} we can form the likelihood, $\mathcal{L}\left(\theta, \phi, y\right)$ according to

$$\mathcal{L}\left(\theta, \phi, y\right) = \prod_i f_{Y_i}\left(y_i, \theta_i, \phi_i\right) = \prod_i \exp\left\{\frac{y_i\theta_i - b\left(\theta_i\right)}{\phi/w_i} + c\left(y_i, \phi, w_i\right)\right\}. \tag{12}$$

We want to maximize this expression w.r.t. every parameter β_j. However, since the logarithm is a monotonically increasing function we may consider the logarithm of the likelihood instead, called the log-likelihood function, $\ell\left(\theta, \phi, y\right)$,

$$\begin{aligned}\ell\left(\theta, \phi, y\right) = \log\left(\mathcal{L}\left(\theta, \phi, y\right)\right) &= \sum_i \left(\frac{y_i\theta_i - b\left(\theta_i\right)}{\phi/w_i} + c\left(y_i, \phi, w_i\right)\right) \\ &= \frac{1}{\phi}\sum_i w_i\left(y_i\theta_i - b\left(\theta_i\right)\right) + \sum_i c\left(y_i, \phi, w_i\right).\end{aligned} \tag{13}$$

Introducing the short hand notation $\eta_i = g(\mu_i)$ for the link function and differentiating w.r.t. the parameters β_j we find that

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_j} &= \sum_i \frac{\partial \ell}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j} = \frac{1}{\phi}\sum_i \left(w_i y_i - w_i b'\left(\theta_i\right)\right)\frac{\partial \theta_i}{\partial \beta_j} \\ &= \frac{1}{\phi}\sum_i \left(w_i y_i - w_i b'\left(\theta_i\right)\right)\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}.\end{aligned} \tag{14}$$

Using the relations found for the exponential family, $\mu_i = b'(\theta_i)$ and $\partial\mu_i/\partial\theta_i = b''(\theta_i)$ we obtain

$$\begin{aligned}&\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} = \frac{1}{v(\mu_i)}, \\ &\frac{\partial \mu_i}{\partial \eta_i} = \left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-1} = \frac{1}{g'(\mu_i)}, \\ &\frac{\partial \eta_i}{\partial \beta_j} = x_{ij},\end{aligned}$$

which inserted into Eq. (14) and setting it equal to 0 gives us

$$\frac{\partial \ell}{\partial \beta_j} = \frac{1}{\phi}\sum_i w_i \frac{y_i - \mu_i}{v(\mu_i)g'(\mu_i)} x_{ij} = 0. \tag{15}$$

with which we get the estimates of the parameters β_j of the model.

Going back to our customers, now that we finally have the maximum likelihood estimates of our model which we plug into Eq. (6), we obtain the probability of each cell by using the mean expression for the logit link in Table 3 according to

$$\pi_i = \frac{1}{1 + \exp\left(\sum_{j=0} x_{ij}\beta_j\right)}.$$

This is the starting point of analyzing which customers we must focus more on and what we can do for them.