SF2930 GLM Lecture 1

January 28, 2019

Contents

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 can we find out which customers will stay with us and which will leave us? This will be the main topic of this lecture which introduce the generalized linear model (GLM), and specifically logistic regression, to answer this question together with the concept of the exponential family.

Table 1: Aggregated historical insurance renewal data based on three sales channels and two groups for the price change including the number of customers, number of renewed insurance contracts and renewal rates for each combination of sales channel and price change group.

Cell	Sales channel	Price change	Number of customers	Number of renewed	Renewal rate
	Variable 1	*Variable 2*	w_i	$z_i = \sum_l y_l$	z_i/w_i
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%

2 From additive linear model to GLM form

In order to ease the introduction of GLM and logistic regression we start by describing the example of cusotmers renewing their insurance contracts, or policies, which we will use throughout this lecture.

We wish to predict which cusotmers stay with us and have historical data that includes two explaining variables, sales channel and price change, as well as the response variable

$$y_l = \begin{cases} 1, & \text{if the insurance contract was renewed} \\ 0, & \text{if the insurance contract was not renewed,} \end{cases}$$

where $l = 1, ..., N$ runs over the entire set of insurance contracts, at our disposal. After dividing the price change into two groups, with a breaking point at 10%, we aggregate the data into *cells*, where a unique combination of the sales channel and the price change group constitutes a cell. This aggregated data is presented in Table 1, which also includes the total number of customers, the number which have stayed with us, thus renewed their insurance contracts, and the renewal rates within each cell.

In general, using a linear additive model we would obtain the expected value, μ_{ij}, of the response variable, Y_i, by

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

where γ_0 is a *base level*, i refers to the customer's group w.r.t. explaining variable number 1, j the group of explaining variable 2 etc. and n is the total number of explaining variables.

In our example we have two explaning variables which gives us

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

where i refers to the customer's sales channel group and j refers to the customer's group w.r.t. price change, e.g. γ_{12} corresponds to Face to face (2). Hence, for each cell there is a corresponding equation on the form of Eq. (1),

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

This model is over parametrized, thus, 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 $\gamma_{11} = \gamma_{21} = 0$. Furthermore, we rename 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

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

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 equal to or greater than 10% price change. Renaming the mean renewal rate for cell i to μ_i and introducing zeros according to

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

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

$$\mu_i = \sum_{j=0}^{3} x_{ij}\beta_j, \tag{3}$$

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

$$x_{ij} = \begin{cases} 1, & \text{if } \beta_j \text{ is included in } \mu_i \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Table 2: Comparison between ordinary linear regression and GLM for the probability distributions and the structure. $N(\mu_i, \sigma_i)$ is the Normal distribution with mean μ_i and standard deviation σ_i, $P(\mu_i, \sigma_i)$ is a probability distribution belonging to the exponential family and $g(\mu_i)$ is the link function.

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

The matrix X which has x_{ij} as elements is often called the *design matrix*.

The general version of Eq. (3) reads

$$\mu_i = \sum_j x_{ij}\beta_j. \tag{5}$$

We have now transformed Eq. (1) to the most basic GLM form in Eq. (5), 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 GLM.

3 GLM and logistic regression

In general, linear regression assume that data come from the Normal distribution with the mean related to the predictors. It is easy to see that this is not always the case, e.g. when trying to find the customers that will leave a company, since then there are only two possible outcomes, stay or leave. Or when modeling the number of claims a customer will have in the future which is non-negative. In both cases the the normal distribution is a poor choice. For instance, the support of the repsonse variables do not coincide with that of the normal distribution.

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

$$g(\mu_i) = \sum_j x_{ij}\beta_j, \tag{6}$$

which is the most general form of GLM. The function, g, is called the *link function* and is the key to solving the problem with having other distributions than then Normal. These main differencies are presented in Table 2.

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 in the next section. However, before that we consider the example of modeling which customers that will stay in detail finding a suitable link function.

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 ✓ that it is compatible.

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

3.1 Logistic regression applied to renewal data

Wether or not a customer will renew its policy can be seen as the outcome of a Bernoulli trial according to

$$\begin{aligned} \Pr(y_l = 1) &= \pi_l, \\ \Pr(y_l = 0) &= 1 - \pi_l, \end{aligned} \tag{7}$$

where $y_l = 1$ corresponds to the customer renewing, thus staying, and $y_l = 0$ means that the customer leaves. In Eq. (7) we have changed to the common notation of π_l instead of μ_l for this particular response variable. The expected value is then found by

$$E[y_l] = 1 \cdot \pi_l + 0 \cdot (1 - \pi_l) = \pi_l.$$

The problem with restricting π_l to the interval $[0, 1]$ is solved by using the *logit* link function according to

$$g(\pi_l) = \ln\left(\frac{\pi_l}{1-\pi_l}\right) = \sum_{j=0} x_{lj}\beta_j, \tag{8}$$

where ln is the natural logarithm, with which

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

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

Instead of having y_l as a response variable we may turn to our aggregated data in Table 1 and note that the number of renewed policies within each cell, Z_i, hence the sum of the individual Bernoulli distributed Y_l, follow the binomial distribution according to

$$Z_i = \text{Bin}(w_i, p_i),$$

where $i = 0, ..., 5$ runs over the number of cells, and p_i is the expected value of the renewal rate in cell i. Since the renewal rate given by

$$p_i = \mathrm{E}[Z_i/w_i],$$

it is also restricted to the interval $[0, 1]$ and we still use the logit link according to

$$g(p_i) = \mathrm{logit}(p_i) = \sum_j x_{ij}\beta_j. \tag{9}$$

Now that we have found a suitable link function to the renewal data in Table 1 we return to the general case to find the expected value and variance of a general response variable with a probability distribution in the exponential family.

4 Exponential family

The exponential family of probability distributions is a generalization of the Normal distribution used in linear models. If the probability distribution of a random variable, Y, depends on a single parameter, θ, and can be written on the form

$$f\,(y;\theta) = \exp\left\{a(y)b(\theta) + c(\theta) + d(y)\right\}, \tag{10}$$

where a, b, c and d are known functions, it beolngs to the exponential family.

If $a(y) = y$ the distribution is in standard form which is often referred to as *canonical form*. $b(\theta)$ is assumed to be twice continuously differentiable with invertible first derivative and is called the *natural parameter* of the distribution. For every choice of such a function we find a family of probability distributions, e.g. the ones listed in Table 3. If there are additional parameters in the functions a, b, c and d these are regarded as known, or *nuisance*, parameters.

4.1 Properties of distributions in the exponential family

We are interested in the expected value and the variance of $a(Y)$, and subsequently for Y, which are the same if the distribution is in canonical form. In order to deduce them we will need two properties, which are valid for any probability density function, namely

$$\int f(y;\theta)\,dy = 1, \tag{11}$$

and

$$\frac{d}{d\theta}\int f(y;\theta)\,dy = \frac{d}{d\theta}1 = 0. \tag{12}$$

4.1.1 Expected value

To find $\mathrm{E}[a(Y)]$ for a member of the exponential family we start by differentiating the probability function in Eq. (10) w.r.t. θ and integrating

$$\int \frac{df(y;\theta)}{d\theta}\, dy = \int \left[a(y)b'(\theta) + c'(\theta)\right] f(y;\theta)\, dy,$$

reversing the order of integration and differentiation on the left hand side this is 0 according to Eq. (12) and we find

$$0 = b'(\theta) \int a(y) f(y;\theta)\, dy + c'(\theta) \cdot 1,$$

where we have used Eq. (11) in the second term. Finally, with the definition of expected value, $\mathrm{E}[a(y)] = \int a(y) f(y;\theta) dy$, and rearanging we find that

$$\mathrm{E}[a(Y)] = -c'(\theta)/b'(\theta). \tag{13}$$

4.1.2 Variance

In order to obtain an expression for $\mathrm{Var}[a(Y)]$ we differentate Eq. (10) twice w.r.t. θ and integrate

$$\int \frac{d^2 f(y;\theta)}{d\theta^2}\, dy = \int \left[a(y)b''(\theta) + c''(\theta)\right] f(y;\theta) + \left[a(y)b'(\theta) + c'(\theta)\right]^2 f(y;\theta)\, dy,$$

where we find that the left hand side is 0 by differentiation Eq. (12) w.r.t. θ. For the first term we have that

$$\begin{aligned}\int \left[a(y)b''(\theta) + c''(\theta)\right] f(y;\theta)\, dy &= b''(\theta)\mathrm{E}[a(Y)] + c''(\theta) \\ &= -b''(\theta)\frac{c'(\theta)}{b'(\theta)} + c''(\theta),\end{aligned}$$

by again using the definition of expected value and Eq. (11) in the first step and then Eq. (13) in the second step. And the second term can be rewritten according to

$$\begin{aligned}\int \left[a(y)b'(\theta) + c'(\theta)\right]^2 f(y;\theta)\, dy &= \int \left[b'(\theta)\right]^2 \left(a(y) - \mathrm{E}[a(Y)]\right)^2 f(y;\theta)\, dy \\ &= \left[b'(\theta)\right]^2 \mathrm{Var}[a(Y)],\end{aligned}$$

where we have used Eq. (13) in the first step, and the definition of variance in the second,

$$\mathrm{Var}[a(Y)] = \int \left(a(y) - \mathrm{E}[a(Y)]\right)^2 f(y;\theta)\, dy.$$

Thus, we get that

$$0 = -b''(\theta)\frac{c'(\theta)}{b'(\theta)} + c''(\theta) + \left[b'(\theta)\right]^2 \mathrm{Var}[a(Y)]$$

which rearanged yields

$$\mathrm{Var}[a(Y)] = \frac{b''(\theta)c'(\theta) - b'(\theta)c''(\theta)}{[b'(\theta)]^3}. \tag{14}$$

4.2 Logistic regression

Returning to our example of customers staying or leaving each observation has the Bernoulli distribution with probability density

$$f(y_l; \pi_l) = \pi_l^{y_l}(1-\pi_l)^{1-y_l}, \tag{15}$$

where $l = 1, ..., N$ which we must be able to write on the form in Eq. (10) in order for us to be able to use GLM. Rewriting Eq. (15) as

$$f(y_l; \pi_l) = \exp\{y_l \ln \pi_l - y_l \ln(1-\pi_l) + \ln(1-\pi_l)\},$$

we see that

$$\begin{aligned}
&a(y_l) = y_l,\\
&b(\pi_l) = \ln \pi_l - \ln(1-\pi_l) = \ln(\pi_l/(1-\pi_l)),\\
&c(\pi_l) = \ln(1-\pi_l),\\
&d(y_l) = 0.
\end{aligned}$$

In order to calculate the expected value and variance we need differentiate b and c

$$\begin{aligned}
b'(\pi_l) &= \frac{1}{\pi_l(1-\pi_l)},\\
b''(\pi_l) &= -\frac{1-2\pi_l}{\pi_l^2(1-\pi_l)^2},\\
c'(\pi_l) &= \frac{1}{1-\pi_l},\\
c''(\pi_l) &= -\frac{1}{(1-\pi_l)^2},
\end{aligned}$$

from which we find the expected value by insertion into Eq. (13) according to

$$\mathrm{E}(Y_l) = -\frac{\frac{1}{(1-\pi_l)}}{\frac{1}{\pi_l(1-\pi_l)}} = \pi_l,$$

and the variance by using Eq. (14)

$$\begin{aligned}
\mathrm{Var}[Y_l] &= \frac{-\frac{1-2\pi_l}{\pi_l^2(1-\pi_l)^2}\cdot\frac{1}{1-\pi_l} - \frac{1}{\pi_l(1-\pi_l)}\cdot\left(-\frac{1}{(1-\pi_l)^2}\right)}{\left(\frac{1}{\pi_l(1-\pi_l)}\right)^3}\\
&= \frac{1-2\pi_l+\pi_l}{\pi_l^2(1-\pi_l)^3}\frac{\pi_l^3(1-\pi_l)^3}{1} = \pi_l(1-\pi_l),
\end{aligned}$$

which are the well known expressions for the Bernoulli distribution.

Similarly, the probability desnity for the aggregated data following the binomial distribution

$$f(y_i; p_i) = \binom{w_i}{y_i} p_i^{y_i} (1-p_i)^{w_i - y_i},$$

for each cell, can be written

$$f(y_i; p_i) = \exp\left\{ y_i \ln p_i - y \ln(1-p_i) + w_i \ln(1-p_i) + \ln \binom{w_i}{y_i} \right\},$$

hence, the only differences are that

$$\begin{aligned} c(\pi_i) &= w_i \ln(1-p_i), \\ d(y_i) &= \ln \binom{w_i}{y_i}, \end{aligned}$$

and, thus, it is also part of the exponential family. Finding the expressions for the expected value and variance is left as exercise.

5 Maximum likelihood estimation of β_j

In this section make the following two common assumptions about the independent random variables Y_i

1. the distribution of each Y_i is in the exponential family, on canonical form[1] and depends on a single parameter θ_i, and,
2. the distributions of Y_i are all on the same form, e.g., all are Poisson distributions, presumably with different parameters θ_i.

With these assumptions we can form the likelihood, $\mathcal{L}(\theta, \phi, y)$, from Eq. (10) according to

$$\mathcal{L}(y; \theta) = \prod_i f_{Y_i}(y_i; \theta_i) = \prod_i \exp\{ y_i b(\theta_i) + c(\theta_i) + d(y_i) \}. \tag{16}$$

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, $\ell(y; \theta)$,

$$\begin{aligned} \ell(\theta, y) = \log(\mathcal{L}(y; \theta)) &= \sum_i \{ y_i b(\theta_i) + c(\theta_i) + d(y_i) \} \\ &= \sum_i y_i b(\theta_i) + \sum_i c(\theta_i) + \sum_i d(y_i). \end{aligned} \tag{17}$$

[1] In practice the most frequently used distributions can be written in canonical form.

Differentiating w.r.t. the parameters β_j we find that

$$\frac{\partial \ell}{\partial \beta_j} = \sum_i \frac{\partial \ell}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} = U_j, \tag{18}$$

which are called the *score statistics*. We will consider each factor separetaly. From Eq. (17) we find that

$$\frac{\partial \ell}{\partial \theta_i} = y_i b'(\theta_i) + c'(\theta_i) = b'(\theta_i)\left(y_i - \mu_i\right),$$

where we have used the expected value $\mathrm{E}[Y_i] = \mu_i$ from Eq. (13) in the second step. Differentiating this expected value then gives

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{-c''(\theta_i)}{b'(\theta_i)} + \frac{c'(\theta_i)b''(\theta_i)}{[b'(\theta_i)]^2} = b'(\theta_i)\mathrm{Var}[Y_i],$$

with the relation for the variance for a member of the exponential family given in Eq. (14). With which we find that

$$\frac{\partial \theta_i}{\partial \mu_i} = 1 \Big/ \left(\frac{\partial \mu_i}{\partial \theta_i}\right) = \frac{1}{b'(\theta_i)\mathrm{Var}[Y_i]}.$$

Finally, introducing the *linear predictor* $\eta_i = g(\mu_i) = \sum_j x_{ij}\beta_j$ the last piece is given by

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}.$$

Thus, we find a score statistic which we set equal to zero in order to find the maximum accoring to

$$U_j = \sum_i \left[\frac{(y_i - \mu_i)}{\mathrm{Var(Y_i)}} \, x_{ij} \frac{\partial \mu_i}{\partial \eta_i} \right] = 0. \tag{19}$$

Solving this equation is typically done by using a statistical analysis software such as R or SAS which uses numerical methods, e.g. Newton-Raphson's method iteratively obtaining the maximum likelihood estimates of the parameters β_j.

Going back to our customers, now that we finally have the maximum likelihood estimates of our model which we plug into Eq. (8), we obtain the probability of each cell by using the mean expression for the logit link in Table 3 according to

$$p_i = \frac{1}{1 + \exp\left(-\sum_{j=0}^{5} x_{ij}\beta_j\right)}.$$

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

Appendices

A Exponential dispersion models

Exponential dispersion models (EDM) are sometimes treated instead of the full exponential family due to their structure which we will explore. It is a generalization of the natural exponential family which in turn is a special case of the exponential family discussed in Section 4.

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

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i, \phi, w_i)\right\}, \tag{20}$$

where

- Y_i is the key ratio in cell i which is a ratio between a the outcome of a random and a volume measure, w_i, of that cell,
- w_i is the weight of the cell, in our case the number of cusotmers,
- 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,
- $b(\theta_i)$ 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 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 restrictions 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 use later.

The importance of the cumulant function is seen in

$$\mu_i = E[Y_i] = \frac{db(\theta_i)}{d\theta_i}, \tag{21}$$

and

$$\mathrm{Var}(\mu_i) = \frac{\mathrm{Var}(Y_i)}{\phi/w_i} = \frac{d\mu_i}{d\theta_i} = \frac{d^2b(\theta_i)}{d\theta_i^2}, \tag{22}$$

for members in the EDM. 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. (21) and Eq. (22).

Using the expression for the probability distribution in Eq. (20) we find

$$\begin{aligned} E\left[e^{tY}\right] &= \int e^{ty} f_Y\left(y;\theta,\phi\right) dy \\ &= \int \exp\left(\frac{y(\theta + t\phi/w) - b(\theta)}{\phi/w} + c(y,\phi,w)\right) dy \\ &= \exp\left(\frac{b(\theta + t\phi/w) - b(\theta)}{\phi/w}\right) \\ &\qquad \times \int \exp\left(\frac{y(\theta + t\phi/w) - b(\theta + t\phi/w)}{\phi/w} + c(y,\phi,w)\right) dy, \end{aligned} \tag{23}$$

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. (20) with $\theta \to \theta + t\phi/w$. Thus, in a neighborhood 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

$$\begin{aligned} \Psi'(0) &= b'(\theta) = E\left[Y\right] = \mu, \\ \Psi''(0) &= b''(\theta)\phi/w = \mathrm{Var}(y), \end{aligned}$$

and

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

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

$$f_{Y_i}(y_i) = \exp\left\{\frac{y_i\mu_i - \mu_i^2/2}{\sigma^2/w_i} + c\left(y_i, \phi, w_i\right)\right\} \tag{24}$$

where

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

B Maximum Likelihood Estimation of β_j in EDM

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

$$\mathcal{L}\left(\theta, \phi, y\right) = \prod_i f_{Y_i}\left(y_i, \theta_i, \phi_i\right) = \prod_i \exp\left\{\frac{y_i\theta_i - b\left(\theta_i\right)}{\phi/w_i} + c\left(y_i, \phi, w_i\right)\right\}. \tag{25}$$

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, $\ell\left(\theta, \phi, y\right)$,

$$\begin{aligned}\ell\left(\theta, \phi, y\right) &= \log\left(\mathcal{L}\left(\theta, \phi, y\right)\right) = \sum_i \left(\frac{y_i\theta_i - b\left(\theta_i\right)}{\phi/w_i} + c\left(y_i, \phi, w_i\right)\right) \\ &= \frac{1}{\phi}\sum_i w_i\left(y_i\theta_i - b\left(\theta_i\right)\right) + \sum_i c\left(y_i, \phi, w_i\right).\end{aligned} \tag{26}$$

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

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_j} &= \sum_i \frac{\partial \ell}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j} = \frac{1}{\phi}\sum_i \left(w_i y_i - w_i b'\left(\theta_i\right)\right)\frac{\partial \theta_i}{\partial \beta_j} \\ &= \frac{1}{\phi}\sum_i \left(w_i y_i - w_i b'\left(\theta_i\right)\right)\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}.\end{aligned} \tag{27}$$

Using the relations found for the EDM, $\mu_i = b'(\theta_i)$ 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. (27) and setting it equal to 0 gives us

$$\frac{\partial \ell}{\partial \beta_j} = \frac{1}{\phi} \sum_i w_i \frac{y_i - \mu_i}{v(\mu_i) g'(\mu_i)} x_{ij} = 0. \tag{28}$$

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