SF2930 GLM Lecture 2

January 30, 2019

1 Introduction

A customer visits If.se to buy a car insurance for one year. How can we price this policy? If we know

(1) *Expected number of claims*, and

(2) *Expected average claim cost*

where a *claim* is an accident that is compensated, we can get the

$$\textit{Expected claim cost} = (1) \times (2) = \textit{"Risk"} \tag{1}$$

The price of the policy is then obtained by

$$\textit{Price} = \textit{Risk} \quad (+\textit{Some extra to pay my salary etc.}) \tag{2}$$

In this lecture we will see how to predict (1) *Expected number of claims*, often referred to as claim frequency, and create a tariff, e.g., a pricing formula.

The data we have at hand is shown in Table 1 which we have aggregated in a similar way as in the previous lecture. Here the explaining variables are the driver's age and car weight. When creating an insurance tariff the grouping is essential, we aim to to find as homogeniuos groups as possible still having enough data in them. This is off course important when grouping the data in any type of analysis using GLM. In a tariff analysis a unique combination of the variables is referred to as a *tariff cell*.

In general there is a weight, w_i, and a random variable, X_i, which together form a *key ratio*, $Y_i = X_i/w_i$, for each tariff cell. These will vary depending on wether it is the claim frequency (w – insurance years, X – number of claims) or extedcted average claim cost, ussually called claim severity, (w – number of claims, X – claim cost). It is possible to model the risk directly according to

$$\begin{aligned}
\textit{Risk} &= \textit{Claim frequency} \times \textit{Claim severity} \\
&= \left(\frac{\textit{Number of claims}}{\textit{Insurance years}}\right) \times \left(\frac{\textit{Claim cost}}{\textit{Number of claims}}\right) \\
&= \frac{\textit{Claimcost}}{\textit{Insurance years}},
\end{aligned}$$

Table 1: Aggregated historic insurance claims data with three groups for Variable 1, Drivar's age, and two groups for Variable 2, Car weight.

Cell	Driver's age	Car weight [kg]	Insurance years	Number of claims	Claims frequency
	Variable 1	*Variable 2*	w	X	$Y = X/w$
1	Young (1)	$0 - 1000$ (1)	500	20	4.00%
2	Young (1)	> 1000 (2)	700	40	5.72%
3	Mid (2)	$0 - 1000$ (1)	1200	50	4.17%
4	Mid (2)	> 1000 (2)	1600	60	3.75%
5	Old (3)	$0 - 1000$ (1)	800	30	3.75%
6	Old (3)	> 1000 (2)	900	35	3.89%

this is sometimes called the pure premium. However, we gain understanding of our model treating claim frequency and claim severity seperately since a explaining variable can have an impact only on either one of them. For instance, installing fire alarms typically influence the severity since since the fire can be put out in an early stage but it will not affect the claim frequency.

2 Multiplicative model for claim frequency

Given the general model form of GLM that we found in the last lecture

$$g(\mu_i) = \sum_{j=0} x_{ij}\beta_j, \tag{3}$$

the starting point is to find the correct distribution for our response variable. With the assumptions

(A1) *Policy independance* - For different insurance policies the number of claims $X_1, X_2, ..., X_n$ are independent,

(A2) *Time independence* - For a policy we may divide the time of the insurance contract into different time intervals which are assumed to be independent, and

(A3) *Homogeneity* - Consider two different policies in the same tariff cell, having the same number of insurance years, then the number of claims X_1 and X_2 have the same probability distribution,

one can argue that this is a Poisson process wich implies a Poisson distribution. This comes with some advantages, e.g., the sum of two independent Poisson distributed variables is itself Poisson distributed. In addition the Poisson distribution is part of the exponential family, hence, we can use GLM and apply the same machinery for the maximum likelihood estimates.

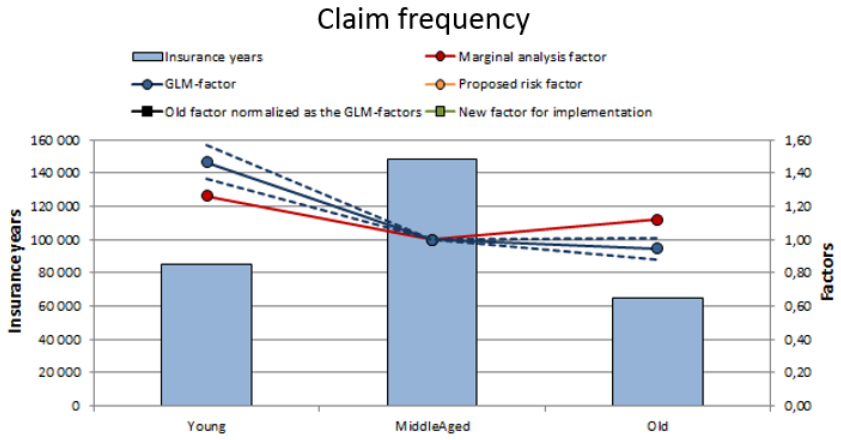

Figure 1: GLM output describing the relative predicted difference in claim frequency for young, middleaged and old drivers.

For the Poisson distribution the log link is the most common choice and it has the great advantage that it will result in a multiplicative model for the mean values since

$$\ln(\mu_i) = \sum_{j=0} x_{ij}\beta_j$$

gives that

$$\mu_i = e^{\sum_{j=0} x_{ij}\beta_j} = e^{x_{i0}\beta_0} \cdot e^{x_{i1}\beta_1} \cdots e^{x_{in}\beta_n},$$

where the dummy varaiables x_{ij} ensure that if β_j is not part of the tariff cell the factor $e^{x_{ij}\beta_j}$ is simply 1 since in that case $x_{ij} = 0$. Furthermore, again considering the funcamental structure of GLM before introducing the dummy variables

$$\begin{array}{lllll}
\mu_1 & = \beta_0 & & & \\
\mu_2 & = \beta_0 & & & +\beta_3 \\
\mu_3 & = \beta_0 & +\beta_1 & & \\
\mu_4 & = \beta_0 & +\beta_1 & & +\beta_3 \\
\mu_5 & = \beta_0 & & +\beta_2 & \\
\mu_6 & = \beta_0 & & +\beta_2 & +\beta_3,
\end{array}$$

we see that the same factor for, e.g., car weight is applied irrespectively of the driver's age. This is a major strength since it makes the pricing more understandable and we may consider one variable at the time, see Figure 1. This also implies that two different customers of different age get the same percentual increase when changing from a light vehicle to a heavy which also adds to the understandability of the model.

In an analysis most of the time is spent on variable selection, feature engineering and grouping the data into homogeneous groups, finnally the parameters

of the actual tariff are set based on the parameter estimates given by the GLM model.

3 Model validation

Having found a model for our claim frequency we want to test it. There are several different ways of testing the model, e.g. wether or not to include an explaining variable, or a group of a specific explaining variable, or how likely is it that we have overfitted the model to the data we have based the model on leading to poor prediction.

3.1 Is every parameter relevant?

A hypothesis test, also known as *Wald test*, can be used to assess specific parameters and be of help in the decision making on wether to include or exclude a specific paramter. A hypothesis test is written as

$$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0 \tag{4}$$

where H_0 is the null hypothesis which may be chosen to indicate that there is no difference from the base cell and H_1 is the alternative hypothesis. However, note that the model corresponding to the null hypothesis need not be a model with only one tariff cell, the only criteria is that the model corresponding to the null hypothesis should be have the same probability distribution, the same link function and be a special case of the model corresponding to the alternative hypothesis.

We know that the estimate of β_j, $\hat{\beta}_j$, is normally distributed, $\hat{\beta}_j \sim N(\beta_j, \sigma^2_{\hat{\beta}_j})$. Estimating the standard deviation of $\hat{\beta}_j$ we can form the test statistic

$$Z_0 = \frac{\hat{\beta}_j}{\hat{\sigma}_{\beta_j}}. \tag{5}$$

If the observation of this test statistic is far enough from 0 then the deviation from the null hypothesis is viewed as significant and safe to use in our model. In other words, we want the confidence interval of β_j

$$I_{\beta_j} : \left[\hat{\beta}_j - 1,96\,\hat{\sigma}_{\beta_j}, \hat{\beta}_j + 1,96\,\hat{\sigma}_{\beta_j}\right], \tag{6}$$

where $1,96 = Z_{\alpha/2}$ with $\alpha = 0,05$, not to overlap with 0, see Figure 2.

Hence, we need to find the *standard error*, the estimate of σ_{β_j}. This is done by taking the following steps.

1. Create the *Hessian matrix*

$$G = \begin{bmatrix} \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_1} & \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_n} \\ \frac{\partial^2 \ell}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 \ell}{\partial \beta_2 \partial \beta_2} & \cdots & \frac{\partial^2 \ell}{\partial \beta_2 \partial \beta_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell}{\partial \beta_n \partial \beta_1} & \frac{\partial^2 \ell}{\partial \beta_n \partial \beta_2} & \cdots & \frac{\partial^2 \ell}{\partial \beta_n \partial \beta_n} \end{bmatrix}. \tag{7}$$

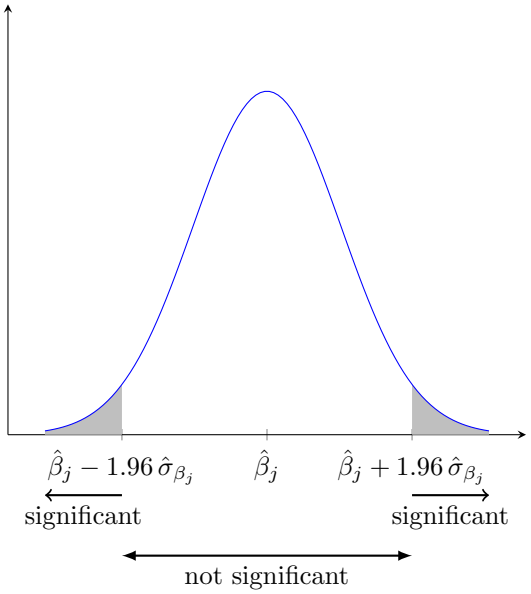

Figure 2: Confidence interval for the test statistic β_j/σ_{β_j} using a significance level of 0.05.

2. Insert the maximum likelihood estimates, $\hat{\beta}_1, \ldots, \hat{\beta}_n$. This gives us actual numbers in the matrix which we call the *evaluated matrix*, $\hat{G}$.

3. Calculate the negative inverse of the evaluated matrix, $-\hat{G}^{-1}$, in which the diagonal element with index (j, j) is $\widehat{\mathrm{Var}}(\hat{\beta}_j)$.

4. The standard error is then $\hat{\sigma}_{\beta_j} = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_j)}$.

3.2 How good does the model fit the data?

Another apsect is how well the model fits the data for which we may use a *likelihood ratio test*.

3.2.1 Likelihood ratio test

This test compares the likelihood of a *full model* (FM) with the likeelihood of a *reduced model* (RM) in the following way

$$\begin{aligned} LR &= 2 \cdot \ln\left(\frac{\mathcal{L}(FM)}{\mathcal{L}(RM)}\right) = 2\left(\ln \mathcal{L}(FM) - \ln \mathcal{L}(RM)\right) \\ &= 2\left(\ell(FM) - \ell(RM)\right), \end{aligned} \tag{8}$$

where ℓ is the log-likelihood.

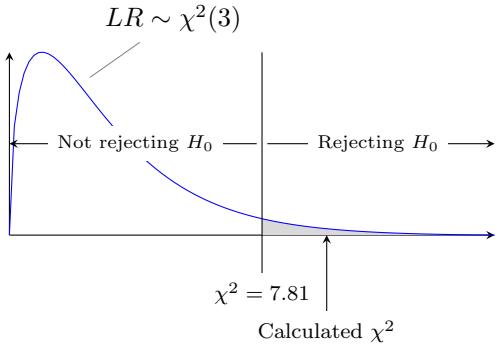

Figure 3: $\chi^2(3)$ distribution of the LR test statistic.

In our example, the FM is our fitted GLM and the reduced model may be a model without any explaining variables

$$\begin{aligned} &\text{FM: } \log(\mu_i) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{in}\beta_n, \\ &\text{RM: } \log(\mu_i) = \beta_0. \end{aligned} \tag{9}$$

Since high log-likelihood value corresponds to a good fit of the data we want LR to be large enough for us to include the explaining variables in the model. Given that we have enough data the LR test will be chi-squared distributed

$$\chi^2 \left(\#\ \textit{parameters in FM - \# parameters in RM} \right). \tag{10}$$

If we compare our model with 4 parameters from driver's age and car weight with the reduced model which uses the same average number of claims for all the customers we find

$$\chi^2(4-1) = \chi^2(3), \tag{11}$$

since we need one parameter in the reduced model to set the average number of claims for the base cell, which then is the only cell. Thus, given that the observation of our LR statistic is larger than some confidence limit, α, we know that the FM is better than the RM, see Figure 3.

If we instead let the FM be the *saturated model* (SM) and our GLM be the reduced model, which we here denote *our model* (OM) we get the definition of the *deviance*

$$D = 2\left(\ell(SM) - \ell(OM)\right). \tag{12}$$

The saturated model is a GLM where we have allowed one parameter β for every observation. This is a perfect model for fitting the data used for the modelling, however, a poor model for predicting the future since this assumes no error or noice at all. This can be compared with an nth grade polynomial perfectly to $n+1$ data points. Hence, the model fits the data perfectly but

fails to capture the trends. This is often called overfitting. Hence, we want the deviance to be as low as possible. See course book p. 430-431[1].

3.2.2 Example: Deviance for a car insurance model of the number of claims

You have created a GLM, which we call the "small model", M_s, to predict the number of insurance claims using the variables *vehicle weight* and *driver's age*. With this model, you get the deviance value D_s.

Now you want to try a new variable, *fuel type*, which has 4 groups

- Petrol (Reference),
- Diesel (β_4),
- Electricity (β_5), and,
- Other (β_6).

One group will be the reference group and is included in the reference tariff cell. Therefor, you add 3 new parameters to the model.

For this new "large model", M_l, you calculate the deviance D_l with which you may compare the two models by calculating

$$\begin{aligned} LR = D_s - D_l &= 2\left(\ell(SM) - \ell(M_s)\right) - 2\left(\ell(SM) - \ell(M_l)\right) \\ &= 2\left(\ell(M_l) - \ell(M_s)\right) \\ &\sim \chi^2\left([\#\text{ parameters in } M_l] - [\#\text{ parameters in } M_s]\right) = \chi^2(3) \end{aligned} \tag{13}$$

If this value is large enough, exceeding α of the $\chi^2(3)$ distribution, the larger model is favorable.

3.2.3 Akaike information criterion

The likelihood test we have used so far will always recommend the larger model if it significantly improves the likelihood by fitting the model data better. However, in real life we often want to keep the model as simple as possible. For example, we only showed that the explaining variable fuel type improves our prediction power. Though, when we sell a car insurance to a customer, is this extra prediction power worth the effort of asking the customer an extra question? However, the concept of keeping the model simple is not solely based on the risk of disturbing a customer with an additional question, in science Occam's razor is utilized for promoting simpler solutions over more complex ones.

The Akaike Information Criterion (AIC) can help us answering this questio. AIC is defined as

$$AIC = 2k - 2\log\left(\hat{\mathcal{L}}\right), \tag{14}$$

[1] D. Montgomery, E. Peck, G. Vining: Introduction to Linear Regression Analysis. Wiley-Interscience, 5th Edition (2012)

where k is the number of parameters in the model (including the intercept β_0) and $\hat{\mathcal{L}}$ is the maximum likelihood (ML) estimate of the GLM. This implies that few parameters and/or high ML value gives low AIC value.

Returning to our example in Sec. 3.2.2 we find the $k_l = 7$ and $k_s = 4$, for the large- and small models, respectively. Thus, if $AIC_{H_l} > AIC_{H_s}$, we might consider excluding fuel type after all. In addition this can also be an indication that the larger model has overfitted the data, which naturally is undesirable in any prediction model.

3.2.4 Bayesian information criterion

If the larger model passes the AIC, $AIC_{H_l} < AIC_{H_s}$ in our example, we may use the Bayesian Information Criterion (BIC) which, in general, punishes additional parameters even more than AIC. BIC is defined by

$$BIC = \log n \cdot k - 2 \log \left(\hat{\mathcal{L}} \right), \tag{15}$$

where n is the number of observations. Tuhs, if $BIC_{H_l} < BIC_{H_s}$ we can feel safe about adding fuel type as variable.