SF2930 GLM Lecture 2

February 20, 2018

1 Introduction

A customer calls If 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* 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 variables are the driver's age and car weight. When creating an insurance tariff the grouping is essential trying to find as homogeniuos groups as possible still having enough data in them. This is off course also the case when working with trying to keep as many customers as possible in the renewals. In a tariff analysis a unique combination of the variables is now referred to as a *tariff cell*.

In general there is a weight w_i, a response variable X_i and a key ratio $Y_i = X_i/w_i$ for each tariff cell. These will vary depending on weather it is the claims frequency (w – insurance years, X – number of claims) or claim severity (w – number of claims, X – Claim cost).

2 Multiplicative Model for Claim Frequency

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

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

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%

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

(A1) *Policy independance* - For different 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.

(A3) *Homogeneity* - Two different policies in the same tariff cell, having the same exposur, 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 apply the same machinery for the maximum likelihood estimates.

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

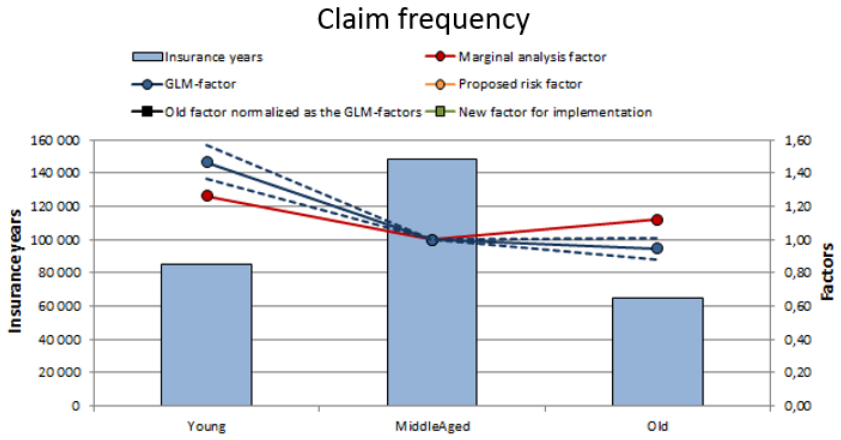

Figure 1: GLM output describing the relative predicted difference in claim frequency for young, middleaged and old drivers.

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 drivar's age. This is a major strength since it makes the pricing more understandable and we may look at one variable at the time, see Figure 1.

3 Model Validation

3.1 Is Every Parameter Relevant?

Now that we have found a model we would like to test it. This can be done througt a hypothesis test, also known as *Wald test*,

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

where H_0 is the null hypothesis. We know that the estimate of β_j, $\hat{\beta}_j$ is normally distributed, $\hat{\beta}_j \sim N(\beta_j, \sigma^2_{\beta_j})$. If we can estimate the standard deviation of $\hat{\beta}_j$ we can form the test statistic

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

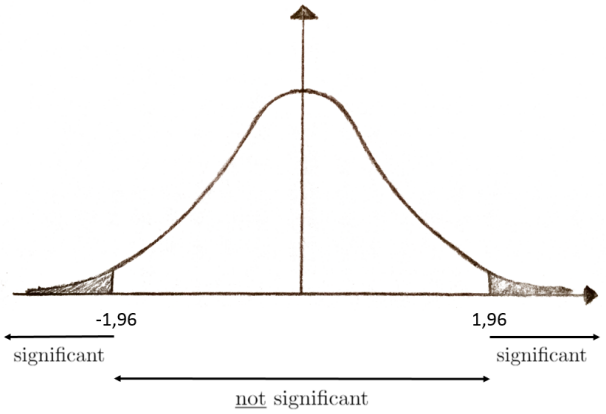

Figure 2: Confidence interval for the test statistic β_j/σ_{β_j} using a significance level of 0.05.

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

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

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*

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

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 $\text{Var}(\hat{\beta}_j)$.

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

3.2 How Good Does the Model Fit the Data?

Now that we have a GLM we would like to know how well it fits the data that we have used to create the model. One way is to do this through 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.

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

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

Which in our example with driver's age and car weight gives us

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

Thus, given that the observation of our LR statistic is larger than some confidence limit, α, we konw 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 uset 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 has an extremely good fit to the data 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].

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

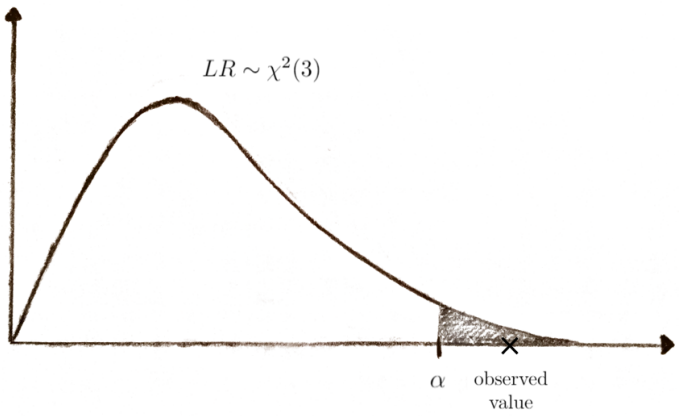

Figure 3: $\chi^2(3)$ distribution of the LR test statistic.

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", H_s, to predict the number of insurance claims using the variables *vehicle weight* and *driver's age*. With this model, you get the deviande 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", H_l, you calculate the deviance D_l with which you may compare the two models by calculating

$$\begin{aligned} LR = D_s - D_l &= 2\left(\ell(SM) - \ell(H_s)\right) - 2\left(\ell(SM) - \ell(H_l)\right) \\ &= 2\left(\ell(H_l) - \ell(H_s)\right) \\ &\sim \chi^2\left(\#parametersH_l - \#parametersH_s\right) = \chi^2(3) \end{aligned} \tag{13}$$

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?

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

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

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.