Generalized Linear Models in Actuarial Science

Silvie Kafková a Lenka Křivánková

2nd–5th May 2013, Malenovice

Outline

Outline

Motivation

- Actuarial science deals with the assessment of risk in insurance and finance.
- Linear regression is insufficient in actuarial mathematics for its strict assumptions.
- A normally distributed random variable does not describe the situation adequately.

Motivation

- This problem is solved by using general linear models.
- GLMs allow to use a distribution different from the normal.
- Any distribution from the exponential dispersion family can be used.
- GLM enables to convert a multiplicative model to an additive model by logarithmic transformation.

Outline

Linear Models

Definition

Given a vector of random variables $\boldsymbol{Y}$, the *linear model (LM)* is

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon,$$

where

- $\varepsilon \sim \mathrm{N_n}(\boldsymbol{0}, \sigma^2 \boldsymbol{I_n})$ is a vector of random variables representing errors,
- β is a vector of the regression parameters,
- $\boldsymbol{X}$ is design matrix (matrix of explanatory variables).

Characteristics of Linear Models

- In a standard linear model we assume normally distributed observations.
- A mean of the observations is a linear function of parameters and explanatory variables.
- The linear models assume independence of variance and mean.
- The mean is linear in explanatory variables.

Real data does not comply to the above mentioned properties of LM. These problems can be solved by working with Generalized linear models instead of ordinary linear models.

Outline

Generalized Linear Models (GLM)

Definition

Given a random variable Y, the *generalized linear model (GLM)* is

$$f(y) = c(y, \phi) \exp\left\{ \frac{y\theta - a(\theta)}{\phi} \right\}, \qquad g(\mu) = \boldsymbol{x}'\boldsymbol{\beta},$$

where

- $f(y)$ is propability density of random variable Y from the exponential family,
- $g(\mu)$ is called the *link function*, it determines relationship between the mean and the explanatory variables x_i.

Outline

Link Function

Definition

The *link function* $g(\mu)$ is a monotonic differentiable function, where

$$g(\mu) = \boldsymbol{x}'\boldsymbol{\beta}.$$

- The link function $g(\mu)$ determines how the mean is related to the explanatory variables x.

Commonly Used Link Functions

Distribution	Link function	$g(\mu)$
normal	identity	μ
poisson	log	$\ln(\mu)$
binomial	logit	$\ln(\frac{\mu}{1-\mu})$
	cloglog	$\ln(-\ln(1-\frac{\mu}{n}))$
exponencial	log	$\ln(\mu)$

Comparing Different Models

- We use the principle of simplicity when comparing the models.
- The simpler model well describing data gets priority over the more complex model that describes the data almost perfectly.

Metods of comparing:

- analysis of deviance
- information criterions

Outline

Analysis of Deviance

With a basic generalized linear model we take into account also the consequent partial models, which are called submodels.

Definition

The *full model*, denoted as GLM_{max}, satisfies the following conditions

- it has the same distribution as the proposed model,
- it has the same link function as the proposed model,
- the number of parameters is equal to the number of the response variables.

Analysis of Deviance

Definition

The *null model*, denoted as GLM_{min}, satisfies the following conditions

- it has the same distribution as the proposed model,
- it has the same link function as the proposed model,
- the number of parameters is equal to one.

The full model- "best regression"
The null model - "worst regression"
The supposed model - between these two extreme models

Analysis of Deviance

Consider GLM with design matrix $\boldsymbol{X}_{n\times m}$ and vector of parameters β_{m}.

Definition

The *submodel*, denoted as GLM_{sub}, with design matrix $\boldsymbol{Q}_{n\times q}$ and vector of parameters β_{q} satisfies the following conditions

- it has the same distribution as the proposed GLM,
- it has the same link function as the proposed GLM,
- the number of parameters is $q < m$ and columns of the design matrix $\boldsymbol{Q}_{n\times q}$ are linear combinations of columns of the design matrix $\boldsymbol{X}_{n\times m}$.

Analysis of Deviance

Definition

The *deviance*, denoted as *dev*, is given by

$$dev = 2(l_{max} - l),$$

where l_{max} is log-likelihood function of the full model and l is log-likelihood function of the proposed submodel.

Analysis of Deviance

Theorem

Consider GLM with vector of parameters $\beta_{\boldsymbol{m}}$ *and its submodel* GLM_{sub} *with* $\beta_{\boldsymbol{q}}$, *where* $q < m < n$.
$\boldsymbol{Y}$ *follows GLM.*
If submodel GLM_{sub} *is suitable then difference of deviances*

$$\Delta dev = dev_{sub} - dev$$

has χ^2 *distribution with degrees of freedom* $(m - q)$.

- If $\Delta dev > \chi^2_{1-\alpha}(m - q)$ then submodel is not suitable and we choose past model.

Outline

Information Criterions

The information criteria indicate a relative measure of the information lost when a model is used to describe reality.

Information criteria for selecting the minimal correct model are

- Akaike information criterion (AIC)
- Schwarz-Bayesian information criterion (BIC)
- Hannan-Quinn information criterion
- Deviance information criterion (DIC)

Akaike Information Criterion

- balance between goodness of fitting data and complexity of the model

Definition

Akaike information criterion (AIC) is given as

$$AIC = -2l + 2k,$$

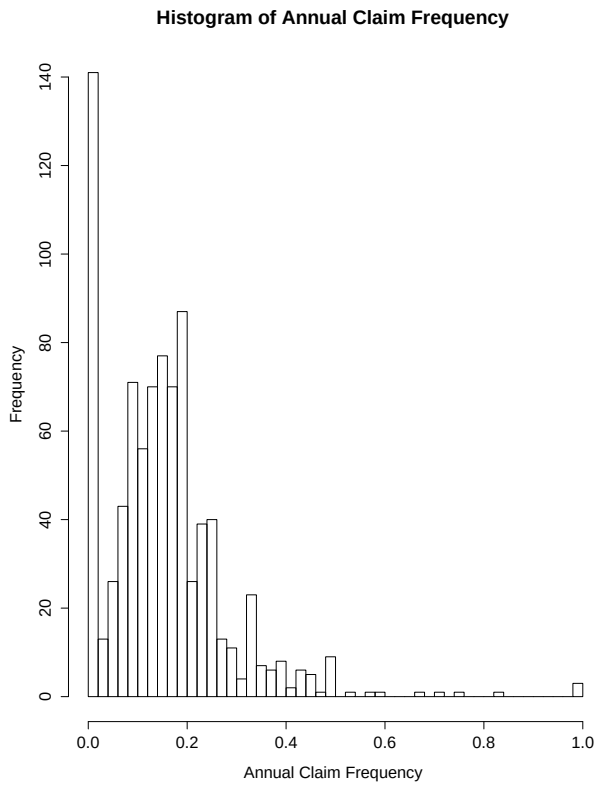
where k is number of parameters and l is log-likelihood function.

- the preferred model is that with the lowest AIC

Outline

Application GLM in Actuarial Science

- we have data set from vehicle insurance
- our aim is predict relation of annual claim frequency on given risk factors
- GLM are suited to the analysis of non-normal data as insurance data



Outline

Insurance Data

- data set is based on one-year vehicle insurance policies taken out in 2004 or 2005
- there are 57 410 policies
- 3 913 policies (6,82%) had at least one claim
- total amount of number of claims is 4 176

Variables in Data Set

Notation	Name of Variable	Range
expo	Exposure	0–1
clm	Claim occurrence	0 (no), 1 (yes)
numclaims	Number of claims	0, 1, 2, 3, 4
veh_body	Vehicle body type	hatchback, sedan, station wagon
veh_age	Vehicle age	1 (new), 2, 3, 4
area	Area of residence	A, B, C, D, E, F
gender	Gender	male, female
agecat	Age band of policy holder	1 (youngest), 2, 3, 4, 5, 6

Outline

Case study

- the drivers can be divided into groups on the basis of the risk factors (gender, age category, area, vehicle body, vehicle age)
- from five risk factors and their options we get 864 groups
- for each of the groups the total amount of exposure during the year is known (expo)
- and observed total of claims (numclaims)
- we model the average number of claims per contract (numclaims/expo)

Tabel of Claim Frequency - example

veh_body	HBACK	gender	F						M					
veh_age	area	agecat	1	2	3	4	5	6	1	2	3	4	5	6
1	A		9.7	23.6	14.8	17.5	10.3	6.3	24.4	7.1	17.5	19.3	16.0	15.8
	B		23.2	13.7	18.0	19.0	9.5	36.7	34.9	10.1	11.5	19.7	11.3	23.3
	C		20.9	15.0	13.4	12.5	6.8	5.5	23.0	17.1	12.6	11.6	12.1	19.4
	D		5.4	43.7	29.1	3.8	13.3	12.3	41.4	0.0	7.8	0.0	15.5	0.0
	E		22.2	8.4	15.9	0.0	28.8	0.0	29.0	0.0	0.0	9.1	16.4	19.4
	F		40.0	0.0	0.0	0.0	0.0	117.4	0.0	0.0	36.1	0.0	0.0	0.0
2	A		20.9	18.8	19.9	13.2	8.7	7.9	24.1	10.3	11.5	17.6	18.9	20.6
	B		25.2	25.4	18.4	14.0	11.1	11.6	6.9	16.5	8.9	17.0	25.6	14.3
	C		19.4	13.5	21.8	17.4	12.4	15.3	25.5	20.1	13.1	8.9	6.8	5.0
	D		11.8	34.4	17.6	7.9	9.8	16.1	25.7	39.7	24.0	5.9	12.5	21.2
	E		16.7	25.7	19.6	18.7	0.0	19.7	0.0	121.1	0.0	0.0	14.4	10.2
	F		36.6	37.6	0.0	14.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	A		13.9	9.9	13.9	13.8	12.5	14.2	31.7	19.5	9.4	15.5	18.6	10.9
	B		20.2	13.9	21.7	11.0	6.4	7.7	45.3	11.6	6.9	13.8	12.9	22.0
	C		20.2	15.3	19.0	20.4	7.0	10.0	21.9	12.2	17.9	15.5	6.4	7.6
	D		3.8	16.9	20.2	6.5	22.8	16.1	0.0	0.0	7.8	7.2	0.0	9.8
	E		55.5	11.8	8.9	18.1	0.0	0.0	0.0	0.0	17.1	8.0	22.7	0.0
	F		11.8	0.0	43.2	0.0	23.3	0.0	0.0	0.0	0.0	45.9	0.0	0.0
4	A		14.7	10.8	13.1	11.9	7.8	5.9	70.6	16.7	12.5	24.1	17.3	4.3
	B		27.3	22.9	12.5	17.3	13.6	22.3	0.0	24.7	15.8	21.9	12.6	3.8
	C		21.3	12.7	19.4	15.2	14.3	9.9	24.1	5.1	20.1	13.4	12.5	10.0
	D		21.5	27.3	14.1	28.5	5.3	5.8	0.0	19.0	22.3	8.5	10.3	0.0
	E		32.9	0.0	19.1	13.8	16.3	6.0	0.0	0.0	0.0	15.5	0.0	0.0
	F		54.9	34.7	24.2	0.0	31.4	0.0	0.0	39.8	32.9	0.0	0.0	0.0

Case study

- we try to find well-fitting GLM for the claim frequency in terms of the risk factors
- for the number of claims on a contract it is reasonable to assume a Poisson distribution or Binomial distribution
- our first GLM for fit data is model with mean-variance relation of the poisson type and log-link
- for creation of the model software R is used

Prediction of parameters

GLM with poisson family and log-link

Coefficients:

(Intercept)	veh_body2	veh_body3	veh_age2	veh_age3	veh_age4
-1.62140	0.06056	0.09926	0.05294	-0.07510	-0.12863
area2	area3	area4	area5	area6	gender2
0.04384	0.01784	-0.11371	-0.01835	0.11873	-0.01132
agecat2	agecat3	agecat4	agecat5	agecat6	
-0.16865	-0.23317	-0.23977	-0.45026	-0.45845	

- the coefficients are given relatively to standard class (veh_body1, veh_age1, area1, gender1, agecat1)
- the coefficients are taken to be zero for the standard class

Prediction of parameters

- According to predicted parameters the best group is the one with veh_body1, veh_age4, area4, gender2 and agecat6.
- The corresponding average number of claims equals

$$e^{(-1.62140-0.12863-0.11371-0.01132-0.45845)} = 0.097$$

- that means one claim each 10.3 years on average

Analysis of deviance

- in the following table we test null hypothesis that adding risk factors actually has no effect
- deviance (dev) for assessing the suitability of the proposed submodel is used
- the difference in deviance (Δdev) between the null model and the proposed model has χ^2 distribution

Analysis of deviance table

GLM with poisson family and log-link

Model specification	df	dev	Δ dev	Δ df
1	855	1048.0		
1+*veh_body*	853	1043.3	4.71	2
1+*veh_age*	852	1027.0	20.99	3
1+*area*	850	1033.2	14.82	5
1+*gender*	854	1047.7	0.38	1
1+*agecat*	850	972.7	75.33	5
1+agecat+*veh_age*	847	953.5	19.16	3
1+agecat+veh_age+*area*	842	942.9	10.58	5
1+agecat+veh_age+*agecat:veh_age*	832	941.1	12.42	15

Analysis of deviance

- according to analysis of deviance the best model is 1+agecat+veh_age
- however we choose model 1+agecat+veh_age+area, although $\Delta dev = 10.58 < \chi^2_{0.95}(5) = 11.07$
- test statistic is close to critical value of χ^2 distribution
- inclusion of the parameter *area* we can support by calculation AIC

Comparing Models by AIC

- Although the AIC penalizes the number of parameters, selected model has smaller AIC than its submodel.
- for model 1+agecat+*veh_age* AIC = 127 900
- for model 1+agecat+veh_age+*area* AIC = 127 200
- hence according to AIC the model is improved
- from experience it is known that factor area is important

Reference

- DENUIT Michel, et al.:
  Actuarial modelling of claim counts: risk classification, credibility and bonus-malus systems
  Hoboken: Wiley. 2007.
- DOBSON Annette J.:
  An Introduction to Generalized Linear Models
  Boca Raton: CRC Press. 2002.
- HELLER Gillian Z., JONG Piet:
  Generalized Linear Models for Insurance Data
  New York: Cambridge University Press. 2008.
- KAAS Rob:
  Modern Actuarial Risk Theory: Using R.
  Heidelberg: Springer. 2009.

Thank you for your attention.