Generalized Linear Models with Actuarial and Financial Applications

Dr. Salvatore Scognamiglio

Università di Napoli "Parthenope"

Lezioni di Tecnica Attuariale delle Assicurazioni

Insurance data Modeling

Models connect variables, and the art of connecting variables requires an understanding of the nature of the variables. Variables come in different forms: discrete or continuous, nominal, ordinal, categorical, and so on.

A model is only as good as the data underlying it. Consequently a good understanding of the data is an essential starting point for modeling. A significant amount of time is spent on cleaning and exploring the data.

It is important to distinguish between different types of variables, as the way that they can reasonably enter a model depends on their type.

Insurance data Modeling

Insurance data is usually organized in a two-way array according to cases and variables. Variables can be quantitative or qualitative. Some example are:

- **Claim amount** is an example of what is commonly regarded as continuous.
- **Legal representation** is a categorical variable with two levels 'no' or 'yes.'
- **Injury code** is a categorical variable, also called qualitative.
- The distribution of **settlement delay** is in the final panel. This is another example of a continuous variable, which in practical terms is confined to an integer number of months or days.

Insurance data Modeling

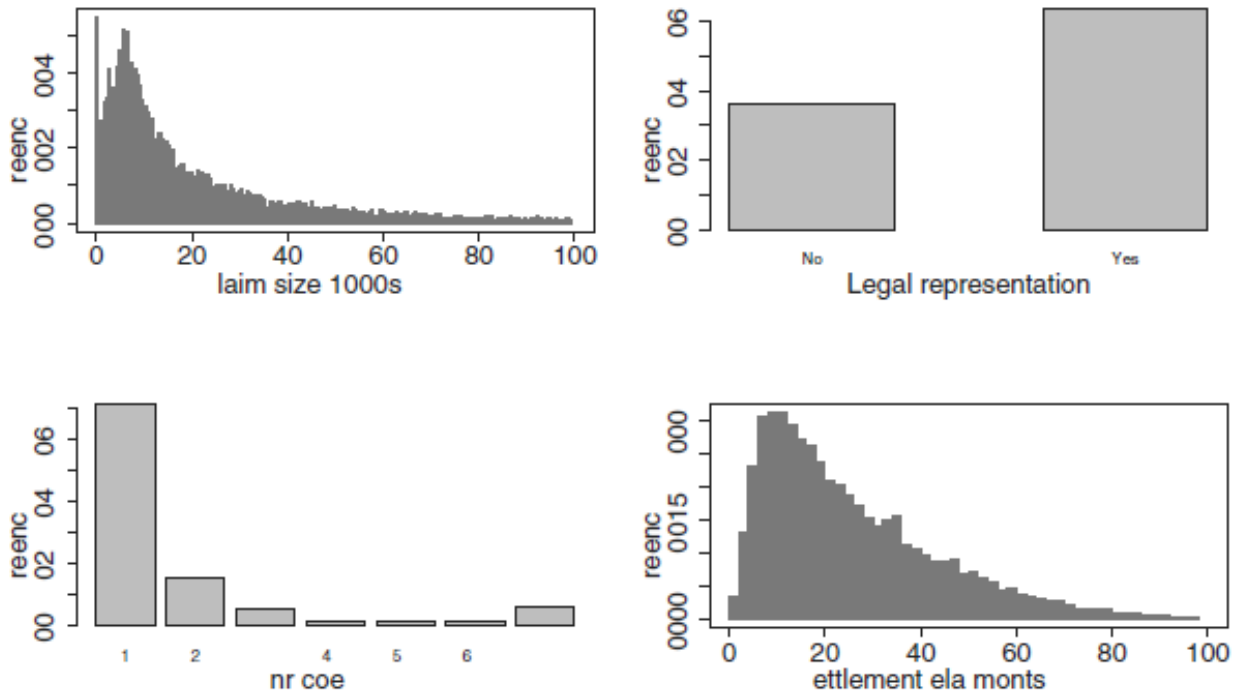


Figure: Graphical representation of personal injury insurance data.

Insurance data Modeling

Sometimes a better understanding of the data and the relationship among the variables can be obtained using mathematical transformations:

- **Histogram of log claim size.** The top left panel displays the histogram of log claim size. Compared to the histogram of actual claim size, the logarithm is roughly symmetric and indeed almost normal.
- **Claim size versus settlement delay.** The top right panel does not reveal a clear picture of the relationship between claim sizes and settlement delay.
- **Claim size versus operational time.** The bottom left panel displays claim size versus the percentile rank of the settlement delay.Note that both the mean and variability of claim size appear to increase with operational time.
- **Log claim size versus operational time.** Log claim size increases virtually linearly with operational time. The log transform has stabilized the variance.

Insurance data Modeling

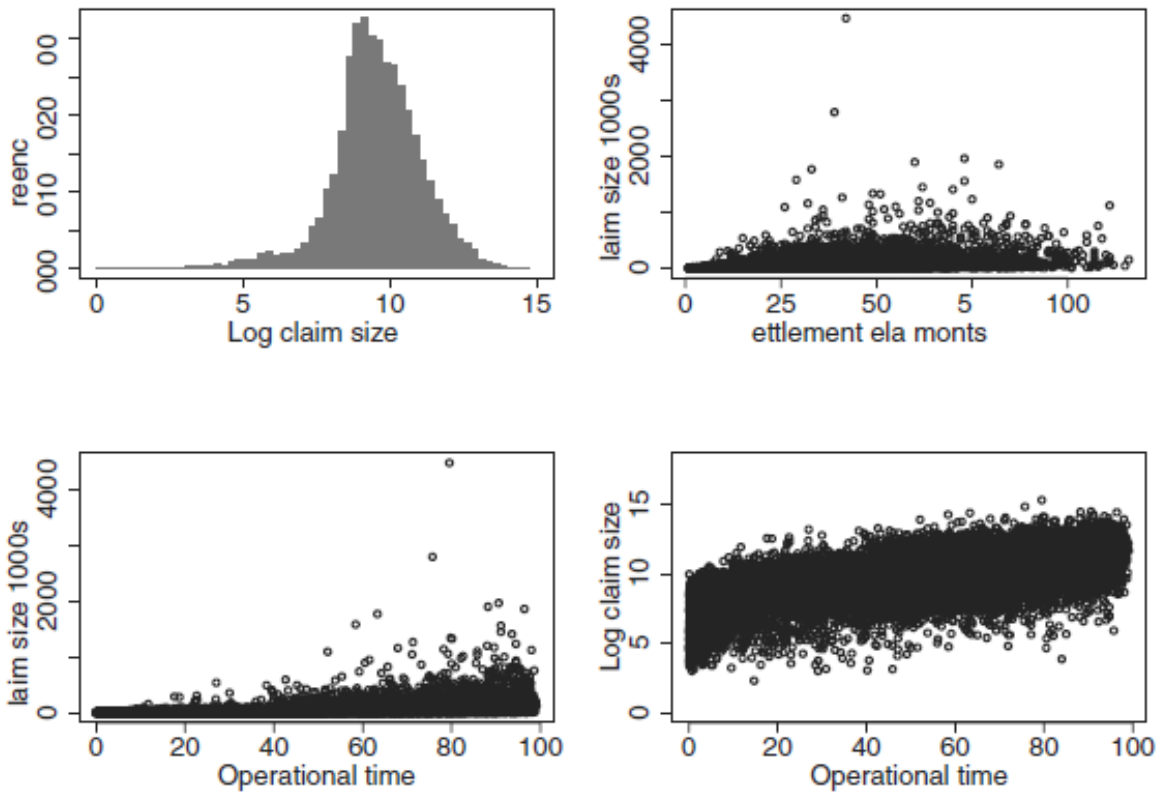


Figure: Relationships between variables in personal injury insurance data set.

Insurance data Modeling

The aim of transformations is to make variables more easily amenable to model, and to tease out trends and effects. Commonly used transformations include:

- **Logarithms.** The log transform applies to positive variables.
- **Powers.** The power transform of a variable y is y^p.
- **Percentile ranks and quantiles**. The percentile rank of a case is the percentage of cases having a value less than the given case.
- **z-score.** Given a variable y, the z-score of a case is the number of standard deviations the value of y for the given case is away from the mean.
- **Logits.** If y is between 0 and 1 then the logit of y is $\log y/(1-y)$. Logits lie between minus and plus infinity, and are used to transform a variable in the $(0, 1)$ interval to one over the whole real line.

Insurance data Modeling

Data exploration using appropriate graphical displays and tabulations is a first step in model building.

It makes for an overall understanding of relationships between variables, and it permits basic checks of the validity and appropriateness of individual data values, the likely direction of relationships and the likely size of model parameters.

Data exploration is also used to examine:

- relationships between the response and potential explanatory variables;
- relationships between potential explanatory variables.

Data displays differ fundamentally, depending on whether the variables are continuous or categorical.

Continuous by continuous

The relationship between two continuous variables is explored with a scatterplot. A scatterplot is sometimes enhanced with the inclusion of a third, categorical, variable using color and/or different symbols.

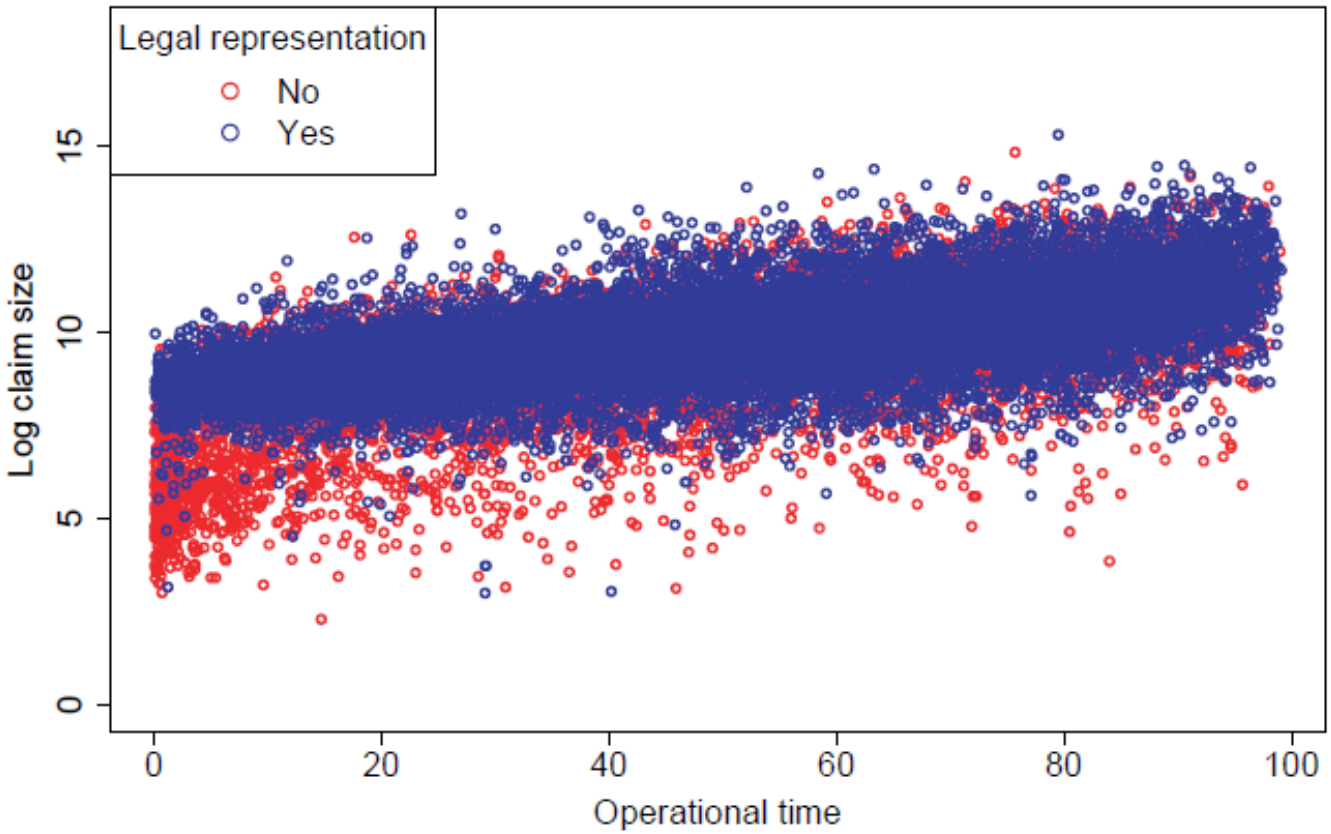


Figure: Scatterplot for personal injury data.

Continuous by continuous

Scatterplot smoothers are useful for uncovering relationships between variables. These are similar in spirit to weighted moving average curves, albeit more sophisticated. Splines are commonly used scatterplot smoothers.

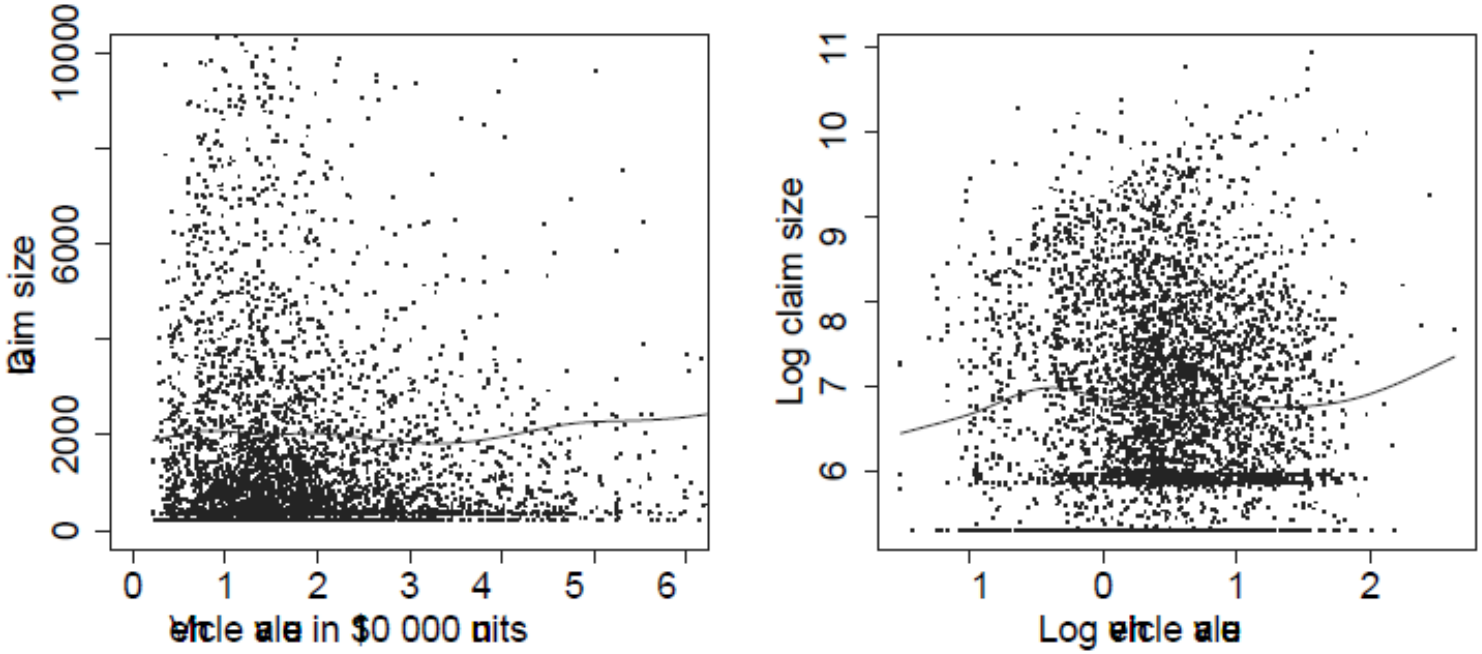


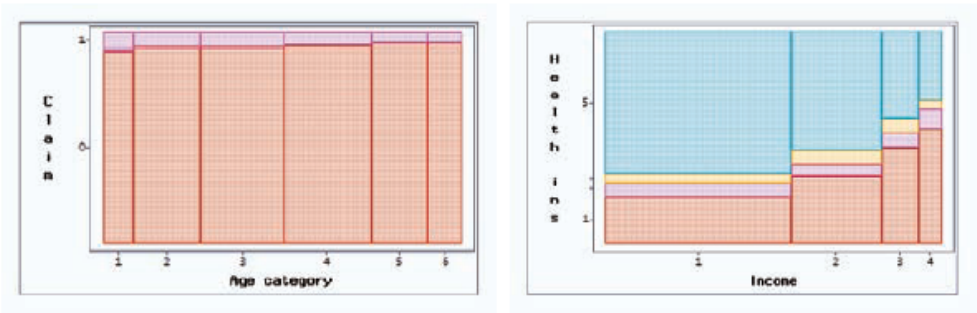
Figure: Scatterplots with splines for vehicle insurance data.

Categorical by categorical

The relationship between two categorical variables can be explored with a frequency table or a Mosaic plots.

Claim by driver's age in vehicle insurance

Claim	Driver's age category 1	2	3	4	5	6	Total
Yes	496	932	1 113	1 104	614	365	4 624
	8.6%	7.2%	7.1%	6.8%	5.7%	5.6%	6.8%
No	5 246	11 943	14 654	15 085	10 122	6 182	63 232
	91.4%	92.8%	92.9%	93.2%	94.3%	94.4%	93.2%
Total	5 742	12 875	15 767	16 189	10 736	6 547	67 856



Vehicle insurance | Private health insurance

Figure: Mosaic plots

Continuous by categorical

Boxplots are appropriate for examining a continuous variable against a categorical variable.

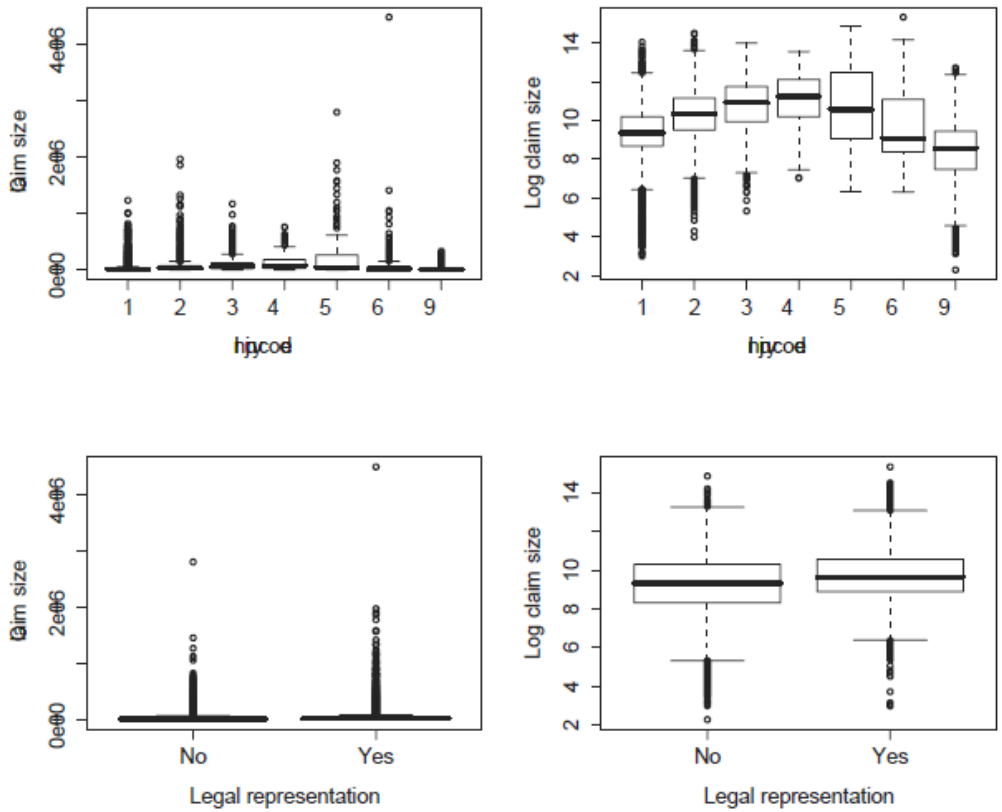


Figure: Personal injury claim sizes by injury code and legal representation.

Assessing distributions

Statistical modeling, including generalized linear modeling, usually makes assumptions about the random process generating the data. For example it may be assumed that the logarithm of a variable is approximately normally distributed.
Distributional assumptions are checked by comparing empirical percentile ranks to those computed on the basis of the assumed distribution.

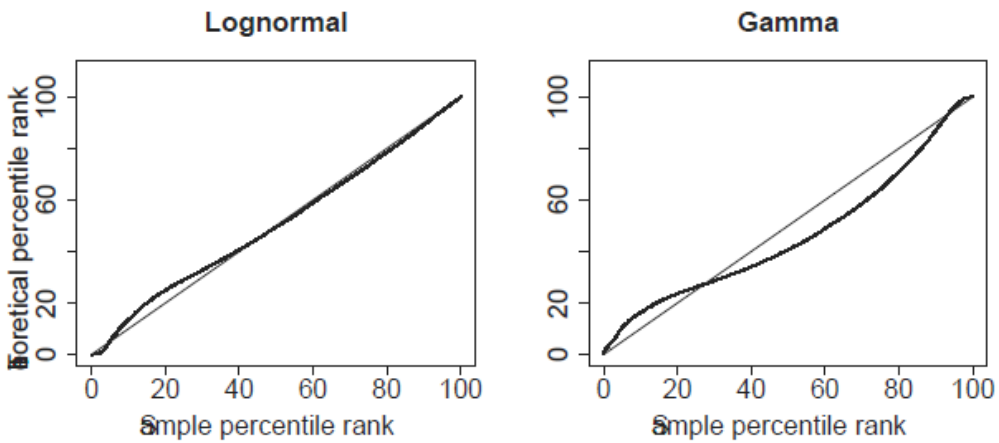


Figure: pp-plots for personal injury claim sizes.

Generalized Linear Model

Multiple linear regression models response variable y_i with $i = 1, \ldots, n$, as a linear function of predictor variables x_{ij}, often called explanatory variables, plus a constant β_0:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

The $k \in \mathbb{N}$ predictor variables are given, nonrandom variables whose values can change with i.
The error term ϵ_i are the differences between the response variables and their predicted values:

$$\epsilon_i = y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}).$$

Two key assumptions are that error terms ϵ_i have expected value 0, $\mathbb{E}[\epsilon_i] = 0$, and the variance of ϵ_i is constant and does not change across observations i, a property referred to as homoskedasticity: $Var[\epsilon] = \sigma^2$. The error term ϵ_i are usually assumed to be independent and normally distributed.

Generalized Linear Model

Taking the variance of both sides $Var[Y_i] = Var[\epsilon_i] = \sigma^2$. The normality of error terms implies that response variable Y_i are normally distributed about their respective means $\mathbb{E}[Y_i]$.

The coefficients in the linear model are estimated by the method of least squares. Normality is not a requirement to construct a linear model using least squares, but is important for hypothesis testing and constructing confidence intervals. Linear models have shown their value in modeling, but in many situations linear models need to be generalized as demonstrated in the following.

Generalized Linear Model

Suppose that Y_i represents the number of claims for risk i in a portfolio of n risks. The actuary may want to predict the expected number of claims for each risk i, $\mathbb{E}[Y_i]$, based on k risk characteristics $x_{i1}, x_{i2}, \ldots, x_{ik}$.
Multiple linear regression may not be the best tool for this job.
Here are three problems with applying the standard linear model:

- The Poisson is commonly used to model the number of claims.
- When modeling the expected number of claims, the left-hand side of equation needs to be non-negative, but this cannot be guaranteed in the linear model.
- Rather than building an additive model where the contributions of risk characteristics $x_{i1}, x_{i2}, \ldots, x_{ik}$ are added, perhaps a multiplicative model is more appropriate.

Some of the complications arising from predicting the number of claims with linear models can be addressed by moving on to generalized linear models.

Generalized Linear Model

Generalized linear models (GLMs) generalize linear regression in two important ways:

- The independent response variables Y_i can be linked to a linear function of predictor variables x_{ij} with a nonlinear link function.
- The variance in the response variables Y_i is not required to be constant across risks, but can be a function of Y_i's expected value.

The GLM predictive equation for response random variables Y_i is

$$g(\mathbb{E}[Y_i]) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

The link function $g(\cdot)$ can be a nonlinear function. In classic linear regression, $g(\cdot)$ is just the identity function $g(x) = x$. There is a restriction on link function $g(\cdot)$ that it be differentiable and strictly monotonic.

Generalized Linear Model

Because it is strictly monotonic, its inverse function exists, and the previous equation can be rewritten as

$$\mathbb{E}[Y_i] = g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$$

The predictor variables x_{ij} are still combined into a linear function, but the response variable $\mathbb{E}[Y_i]$ can be a nonlinear function of this linear combination. The linear function of predictor variables is often assigned the symbol η:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}.$$

Letting $\mu_i = \mathbb{E}[Y_i]$ yields a shorthand equation

$$\mu_i = g^{-1}(\eta_i).$$

The other important GLM assumption is that random variables Y_i can be members of a linear **exponential family of distributions**.

Exponential Family of Distributions

For GLMs, response variable is assumed to have a probability distribution function that can be written as:

$$f(y;\theta,\phi) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right].$$

Note that function $c(y,\phi)$ does not include parameter θ.
Parameter θ is often refereed to as the **canonical parameter**, or **natural parameter**, or parameter of interest.
Parameter ϕ is called the **dispersion parameter** or, sometimes, nuisance parameter because the mean of the distributon does not depend directly on ϕ.
The function $b(\theta), a(\phi)$ and $c(y,\phi)$ determine the type of distribution.

Exponential Family of Distributions

The mean and variance of the distribution are simply

$$\mathbb{E}[Y] = b'(\theta),$$

$$Var[Y] = a(\phi)b''(\theta),$$

where $b'(\theta)$ is the first derivative with respect to θ and $b''(\theta)$ is the second derivative.
Distributions in this exponential family include the normal, binomial, Poisson, exponential, gamma, inverse-Gaussian and the compound Poisson-gamma.
With a little algebra the common forms of these distribution can be rewritten in exponential family form.

The Poisson Distribution

The Poisson distribution is a member of the exponential family with probability mass function

$$\begin{aligned} f(y;\lambda) &= \frac{\lambda^y e^{-\lambda}}{y!} \\ &= exp\left[\log\left(\frac{\lambda^y e^{-\lambda}}{y!}\right)\right] \\ &= exp\left[\frac{y\log\lambda - \lambda}{1} - \log y!\right] \end{aligned}$$

Making the substitution $\theta = \log\lambda$ produces

$$f(y;\theta) = exp\left[\frac{y\theta - e^{\theta}}{1} - \log y!\right].$$

Note that $b(\theta) = e^{\theta}$ and $c(y,\phi) = -\log y!$. We can let $a(\phi) = \phi = 1$, calculating the mean and the variance of the distribution

$$\mathbb{E}[Y] = b'(\theta) = e^{\theta} = \lambda$$

$$Var[Y] = a(\phi)b''(\theta) = e^{\theta} = \lambda.$$

The Poisson Distribution

If a random variable has the Poisson distribution, **its expected value and variance are equal**.
Real data that might be plausibly modelled by the Poisson distribution often have a larger variance and are said to be **overdispersed**, and the model may have to be adapted to reflect this feature.

The Normal Distribution

Suppose that $Y \sim N(\mu, \sigma^2)$. The normal distribution is a member of the exponential family:

$$\begin{aligned} f(y;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \\ &= exp\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\right] exp\left[-\frac{y^2-2y\mu+\mu^2}{2\sigma^2}\right] \\ &= exp\left[\frac{y\mu-\mu^2/2}{\sigma^2} - \left(\frac{y^2}{2\sigma^2} + \log(\sqrt{2\pi\sigma^2})\right)\right] \end{aligned}$$

The parameter μ corresponds to θ and $\phi = \sigma^2$. Making the substitutions we can rewrite the pdf

$$f(y;\theta,\phi) = exp\left[\frac{y\theta-\theta^2/2}{\phi} - \left(\frac{y}{2\phi} + \log(\sqrt{2\pi\phi})\right)\right].$$

So, $b(\theta) = \theta^2/2$, $a(\phi) = \phi$, and $c(y,\phi) = -(y^2/2\phi^2 + \log(\sqrt{2\pi\phi})$ and

$$\mathbb{E}[Y] = b'(\theta) = \frac{d(\theta^2/2)}{d\theta} = \theta = \mu,$$

$$Var[Y] = a(\phi)b''(\theta) = \phi\frac{d^2(\theta^2/2)}{d\theta^2} = \phi = \sigma^2.$$

The Normal Distribution

The Normal distribution is widely used for three main reasons.

1. First, many naturally occurring phenomena are well described by the Normal distribution; for example, height or blood pressure of people.
2. Second, even if data are not Normally distributed (e.g., if their distribution is skewed) the average or total of a random sample of values will be approximately Normally distributed; this result is proved in the Central Limit Theorem.
3. Third, there is a great deal of statistical theory developed for the Normal distribution, including sampling distributions derived from it and approximations to other distributions.

For these reasons, if continuous data y are not Normally distributed it is often worthwhile trying to identify a transformation, such as $y' = \log y$ or $y' = \sqrt{y}$ which produces data y' that are approximately Normal.

Exponential Family of Distributions

Common form of pdf	θ	$b(\theta)$	ϕ	$a(\phi)$	$c(y, \phi)$
Normal: $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$	μ	$\theta^2/2$	σ^2	ϕ	$-\frac{1}{2}\left[\frac{y^2}{\phi} + \ln(2\pi\phi)\right]$
Poisson: $\lambda^y e^{-\lambda}/y!$	$\ln\lambda$	e^θ	1	1	$-\ln(y!)$
Binomial[a]: $\binom{n}{y} p^y(1-p)^{n-y}$	$\ln[p/(1-p)]$	$n\ln(1+e^\theta)$	1	1	$\ln\binom{n}{y}$
Gamma[b]: $\beta^\alpha y^{\alpha-1} e^{-\beta y}/\Gamma(\alpha)$	$-\frac{\beta}{\alpha}$	$-\ln(-\theta)$	$\frac{1}{\alpha}$	ϕ	$\frac{1}{\phi}\ln\frac{y}{\phi} - \ln y - \Gamma\left(\frac{1}{\phi}\right)$
Inverse-Gaussian: $\sqrt{\frac{\lambda}{2\pi y^3}} \exp\left[\frac{-\lambda(y-\mu)^2}{2y\mu^2}\right]$	$\frac{-1}{2\mu^2}$	$-\sqrt{-2\theta}$	$\frac{1}{\lambda}$	ϕ	$-\frac{1}{2}\left[\ln(2\pi\phi x^3) + \frac{1}{\phi y}\right]$

Notes: Other sources may show different parameterizations. Common forms of pdfs are displayed so that readers can make their own calculations and reconcile formulas.
[a] n is assumed to be a known, fixed quantity.
[b] n With this parameterization the mean and variance are $\mathrm{E}[y] = \alpha/\beta$ and $\mathrm{Var}[y] = \alpha/\beta^2$.

Figure: Exponential Family Form.

The Link Function

The link function must be differentiable and strictly monotonic-either strictly increasing or strictly decreasing - so that its inverse exist:

$$\begin{aligned} g(\mu_i) &= \eta_i, \\ \mu_i &= g^{-1}(\eta_i). \end{aligned}$$

where

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}.$$

The modeler has a choice of link functions, but some links may be more appropriate than others for a model. For example, an important consideration is selecting the link function is the range of $\mu_i = \mathbb{E}[Y_i]$.

Response Variable Y_i is Number of Claims

The expected number of claims μ_i has range $(0, \infty)$. The linear predictor η_i may have range $(-\infty, \infty)$.
It may be that $\eta_i < 0$ for possible combinations of predictors x_{ij}. A solution to this contradiction is a log-link function. The log link is $g(\mu) = \log(\mu)$:

$$\log(\mu) : (0, \infty) \to (-\infty, \infty).$$

The inverse of log link $g^{-1}(\eta) = e^{\eta}$ maps $(-\infty, \infty)$ onto $(0, \infty)$.

μ_i is Probability of an Event

The GLM may model probability of events such as customer renewing policies or claims being fraudulent.
The response variable Y will take on values of 1 or 0 depending on whether the event happens or not and μ_i will be the probability of the event.
Probabilities μ_i have range $[0, 1]$. As discuss earlier, η_i has possible range $(-\infty, \infty)$. An appropriate link function can be costructed in two steps. If p is the probability of an event, then the odds-ratio is $p/(1-p)$:

$$p/(1-p) : (0,1) \to (0, \infty).$$

Next take the log of the odds-ratio:

$$\log(p/(1-p)) : (0,1) \to (-\infty, \infty).$$

Link $g(\mu) = \log(\mu/(1-\mu))$ is called logit link. Two other common link for this mapping are:

- Probit: $g(\mu) = \Phi^{-1}(\mu)$ where $\Phi^{-1}(\cdot)$ is the inverse standard cumulative normal distribution.
- Complementary log-log: $g(\mu) = \log(-\log(1-\mu))$.

Link Functions

	$g(\mu)$	$g^{-1}(\eta)$	Range of $g^{-1}(\eta)$
identity	μ	η	$(-\infty, \infty)$
log	$\ln(\mu)$	e^{η}	$(0, \infty)$
logit	$\ln[\mu/(1-\mu)]$	$e^{\eta}/(1+e^{\eta})$	$(0, 1)$
probit	$\Phi^{-1}(\mu)$	$\Phi(\eta)$	$(0, 1)$
complementary log-log	$\ln(-(\ln(1-\mu))$	$1-e^{-e^{\eta}}$	$(0, 1)$
inverse	$1/\mu$	$1/\eta$	$(-\infty, 0) \cup (0, \infty)$
inverse squared	$1/\mu^2$	$1/\sqrt{\eta}$	$(0, \infty)$

Figure: Some link functions.

Maximum Likelihood Estimation

The coefficients $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)'$ for the GLM are estimated from the data using maximum likelihodd estimation (MLE). Choosing a distribution to model random variables Y_i allows one to apply MLE. The likelihood function is

$$L(\boldsymbol{y}; \boldsymbol{\beta}) = \prod_{i=1}^{n} exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c_i(y_i, \phi)\right].$$

The left-hand side shows that the likelihood is a function of the n observations $y_1, \ldots, y_n$ and the parameters $\boldsymbol{\beta}$. It is easier to maximise the log-likelihood:

$$l(\boldsymbol{y}; \boldsymbol{\beta}) = \sum_{i=1}^{n}\left[\frac{y_i\theta_i + b(\theta_i)}{a_i(\phi)} + c_i(y_i, \phi)\right].$$

The log-likelihood can be maximized by calculating partial derivatives with respect to the β_i's and setting them equal to zero.

$$\frac{\delta l(\boldsymbol{y}; \boldsymbol{\beta})}{\delta \beta_j} = \sum_{i=1}^{n}\frac{\delta}{\delta \beta_j}\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c_i(y_i, \phi)\right]$$

Maximum Likelihood Estimation

Statistical packages have numerical methods to maximize the log-likelihood function. A technique commonly used is referred to as iteratively reweighted least squares. One also sees the name Fisher Scoring algorithm. For those interested in the details of these numerical techniques, see Dobson and Barnett (2008); Gill (2000); McCullagh and Nelder (1997); and Nelder and Wedderburn (1972).

Generalized Linear Model: a summary

Summarizing:

- Response variables Y_i have a distribution from the exponential family and are independently distributed.
- Predictor variables x_{ij} are combined into linear predictors plus a constant

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}.$$

- Link function $g(x)$ is strictly monotonic and differentiable with inverse function $g^{-1}(x)$.
- The expected values of Y_i, $\mu_i = \mathbb{E}[Y_i]$, are predicted by the equations

$$g(\mu_i) = \eta_i \text{ or } \mu_i = g^{-1}(\eta_i) \text{ for } i = 1, \ldots, n.$$

- Coefficient $\beta_0, \beta_1, \ldots, \beta_k$ are estimated from data using maximum likelihood estimation.
- The modeler must choose the distribution and link function appropriate for the model.

Comparing Models

A statistical measure called deviance is commonly used to evaluate and compare GLMs and it is based on log-likelihoods. The log-likelihood function for linear exponential family distributions is

$$l(\boldsymbol{y};\boldsymbol{\theta}) = \sum_{i=1}^{n} \left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c_i(y_i,\phi) \right]$$

When a particular GLM is constructed - let's call it model M - coefficients β_j are calculated to maximize the log-likelihood. The canonical parameters can be computed using these coefficients and predictive variables for the chosen distribution and link function.
Let $l(\boldsymbol{y};\boldsymbol{\theta}^M)$ denote the value of the log-likelihood for these parameters $\hat{\theta}_i^{\,M}$.
One way of assessing the fit of a given model is to compare it to the model with the best possible fit. The best fit will be obtained when there are as many parameters as observations: this is called a saturated model. A saturated model S will ensure there is complete flexibility in fitting $\theta_i = y_i$.

Comparing Models

The difference between the log-likelihoods of the saturated model S and model M is

$$l(\boldsymbol{y};\hat{\boldsymbol{\theta}}^S) - l(\boldsymbol{y};\hat{\boldsymbol{\theta}}^M) = \sum_{i=1}^{n}\left[\frac{y_i(\hat{\theta}_i^S - \hat{\theta}_i^M) - b(\hat{\theta}_i^S) + b(\hat{\theta}_i^M)}{a_i(\phi)}\right].$$

If $a_i(\phi) = \phi$, the Deviance of M from S is defined

$$\begin{aligned} D(\boldsymbol{y};\hat{\boldsymbol{\theta}}^M) &= 2\phi[l(\boldsymbol{y};\hat{\boldsymbol{\theta}}^S) - (\boldsymbol{y};\hat{\boldsymbol{\theta}}^M)] \\ &= 2\sum_{i=1}^{n}[y_i(\hat{\theta}_i^S - \hat{\theta}_i^M) - b(\hat{\theta}_i^S) + b(\hat{\theta}_i^S)]. \end{aligned}$$

We observe that $D(\boldsymbol{y};\hat{\boldsymbol{\theta}}^M) > 0$ and $D(\boldsymbol{y};\hat{\boldsymbol{\theta}}^S) = 0$.

References


Dobson, A. J., Barnett, A. G. (2018).
An introduction to generalized linear models.
CRC press.


Dunn, P. K., Smyth, G. K. (2018).
Generalized linear models with examples in R.
Springer.


Frees, E. W., Derrig, R. A., Meyers, G. (Eds.). (2014).
Predictive modeling applications in actuarial science (Vol. 1).
Cambridge University Press.


Frees, E. W., Meyers, G., Derrig, R. A. (Eds.). (2016).
Predictive Modeling Applications in Actuarial Science: Volume 2, Case Studies in Insurance.
Cambridge University Press.