GLMs in insurance

(technical foundations)

2. Technical foundations

GLMs → model relationship between a variable whose outcome we want to predict and 1+ explanatory variables

called target variable y

y = claim freq. / claim severity / premium / loss ratio

GLM estimates probability - produces expected value

explanatory variable (predictors) : $x_1, x_2, \dots, x_p$

↳ e.g age, car type, marital status

2.1 Components of GLM

target variable outcome driven by systematic component and random component

Systematic component → portion of variation in outcomes related to the predictors. E.g: if driver age effects claim freq. and is in model

Random component → Portion of outcome Not driven by predictors

2.1.1 Random component (exponential family)

target variable modelled as random variable that follows probability dist^{n}

dist^{n} is member of exponential family

↳ includes: normal, Poisson, binomial, gama, Tweedie

Randomness of particular risk $y_i \sim \exp(\mu_i, \phi)$

ϕ = dispersion parameter (like variance)

μ = mean of dist^{n} (expected value)

↳ models output

↳ For GLMs, const!

2.1.2 Systematic component

GLMs model relation between μ_i and predictors as ...

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$

$g(\mu_i)$: Link f^n, transformation of μ

$\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$: Linear predictor

β values predicted by GLM software

Note: we are interested in μ_i, not $g(\mu_i)$

Link f^n gives flexibility "general"

Still linear comb^n of predictors "Linear model"

For insurance plan, $g(x) = \ln(x)$

$$\ln \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})$$

$$= e^{\beta_0} e^{\beta_1 x_{i1}} e^{\beta_2 x_{i2}} \dots e^{\beta_p x_{ip}}$$

Pros ...

① Easy to implement

② additive terms can give -ve premiums (unphysical). Multiplicative terms mean don't need to implement min. premium rules (clunky)

③ % based penalties make more sense than + £100 premium regardless of whether base premium is £1M or £1K vs 10%

2.2 Exponential family variance

$$\text{Var}[y] = \phi \cdot V(\mu)$$

variance = dispersion param. × 'variance fⁿ'

variable fⁿ depends on distⁿ

higher risk = higher variance

2.3 Variable significance

For each predictor, GLM software estimates its coeff.

∴ different data = different coeff's.

If predictor makes no difference to target, coeff = 0

Statistics used to evaluate 'accuracy' of coeff's →

standard error, p-value, confidence int.

2.3.1 Standard error

coeff estimate is result of standard process

small error = more confidence

large data sets generally have smaller error

larger ϕ = larger error, since related to variance

'more noise'

2.3.2 p-value

p-value is an estimate of the probability of a value of certain magnitude arising by pure chance

E.g. model gives coeff. = 1.5 with p-val = 0.0012 -> indicates if coeff is 0, chance of getting 1.5+ is 0.0012

∴ ∵ 1.5 so unlikely, we say odds of coeff being 0 in unlikely -> is statistically significant

If p-val = 0.52, coeff more likely to be 1.5. If var has no effect - we have evidence for it

Null hypothesis = true value of variable is 0. For small p-value - can reject null hypothesis (accept variable has non zero effect on outcome)

2.3.3 Confidence interval

Null hypothesis doesn't have to be 0

What range of values would not be rejected at our chosen p-value ? — this is confidence interval (reasonable range of estimates for coeff)

For p-value = 0.05. GLM software (SAS) will give 95% confidence interval will look like [0.17, 0.79] for coeff = 0.48

2.4 Types of predictor variables

Continuous variable - numeric value that represents a measurement on a continuous scale (age, population density)

categorical variable - takes 1 or 2 or ... possible values (numeric or non-numeric) (Vehicle type -> SUV, sedan. Vehicle use -> commute, pleasure)

2.4.1 Treatment of continuous variables

Continuous variables input into GLM as is

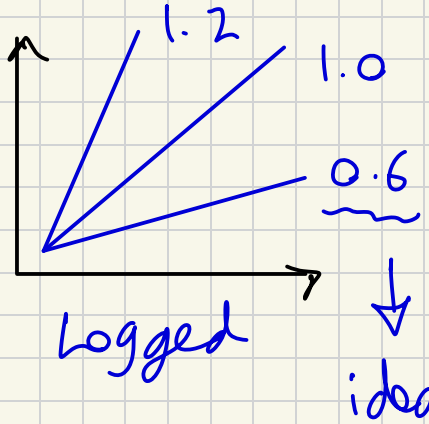
Often appropriate to take ln(·) of predictors before putting them into model so scales match (target vs predictor)

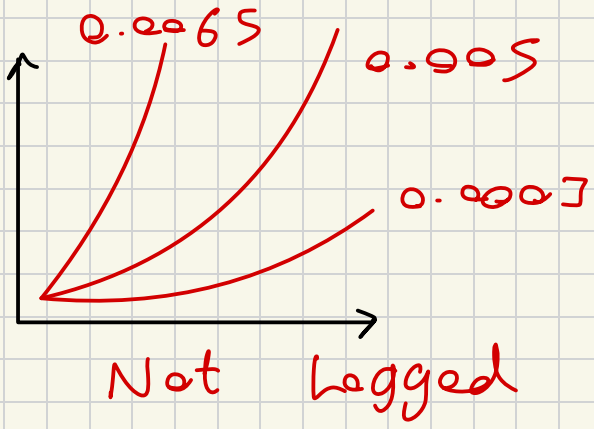
→ $\ln \mu = \beta_0 + \beta_1 \ln x$ ← Now a power transform of original variable

$$\mu = e^{\beta_0} \cdot e^{\beta_1 \ln x} = e^{\beta_0} \cdot x^{\beta_1}$$

E.g. Coeff = 0.62. House worth £200 K will have higher x = AOI than £100 K by

$200^{0.62} / 100^{0.62} = 1.54 = 54\%$ more





Generally considers 'logged form' natural state and original variable 'transformed'

Cannot log -ve values - not always feasible

Assumes Linear relation between x and μ

2.4.2 Treatment of categorical variables

One level is assigned base level 1, all others = 0.

E.g for sedan, truck, SUV. Start with SUV = 1, sedan, truck = 0. Repeat for all vtypes. End up with design matrix

→ $g(\mu) = \beta_0 + 1.23 \cdot 1 + 0.57 \cdot 0 - 0.30 \cdot 0 + \dots$

$\dots + \beta_4 x_4 + \dots + \beta_p x_p$

non-sedans drop out eqn

greatest coeff = higher claim freq.

$g(\mu) = \beta_0 + 1.23 + \beta_4 x_4 + \beta_p x_4$

If use Log-link GLM ... If base level for sedan, $e^0 = 1$. For SUV, $e^{1.23} = 3.42$

∴ expected frequency for SUVs ~ 242% higher than sedan.

2.4.3 Base Levels

Changing base level doesn't affect model predictions

Coeff's shift since relative to a different base

p-values and confidence interval change!

If base has less data, estimates are less reliable

Choose base level as well populated category

2.5 Weights

Sometimes, dataset going into GLM will include rows that represent the averages of the outcomes of individual risks

Rows that represent a greater no. risks should carry more weight

Recall $Var[y] = \phi V[\mu]$ ← exp family variance

With weight factor $Var[y_i] = \frac{\phi V[\mu_i]}{w_i}$

When weight = no. records that an aggregated row represents

For random variable X, $Var[(\Sigma X_i)/n] = \frac{1}{n} Var[X]$

→ variance of n independent and identically distributed random variables = $1/n$ times variance of one such random variable

∴ A row representing AVG loss of two claims would be expected to have half the variance of single-claim row

2.6 Offsets

When modelling insurance rating plans - often updating some elements, not whole plan at once. E.g's ...

Rating algorithms -> base loss cost which varies by region or class - remains static

Deductible factors -> calculated using traditional loss elim - based techniques. GLM used for factors other than deductible

Fixed variable not assigned coeff by GLM - but GLM must be aware of its existence so other variables optimal in its presence.

Done with an offset.

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \underbrace{\text{offset}}_{(\text{coeff} = 1 \text{ by def}^n)}$$

Exposure offsets -> when modelling target variable that varies with time / method of exposure

offset adjustment to mean. Weight is an adjustment to the variance

2.7 Distributions

2.7.1 Distributions for severity

Gamma dist^{n} -> Right skewed, LB = 0, sharp peak, long tail to right.

Gamma variance f^{n} $V(\mu) = \mu^2$ -> assumed variance of severity for claim proportional to exp f^{n} of mean

Inverse Gaussian dist^{n} -> right skewed, LB = 0. Sharper peak, wider tail compared to gamma.

IG variance f^{n} $V(\mu) = \mu^3$ -> also scales exponentially, but faster rate

2.7.1 Distributions for frequency

Poisson dist^{n} → models count of events within a fixed time interval. Typically a discrete dist^{n}, in GLM can take on fractional values – useful when claim count divided by premium or exposure

Variance f^{n} → $V(\mu) = \mu$ (linear)
'True' Poisson, variance = mean ($\phi = 1$).
Generally, variance > mean (overdispersion)
Use QDP so ϕ can take any +ve value so variance not understated

Negative binomial dist^{n} → To deal with OD in Poisson dist^{n} from random variation in Poisson mean is to treat mean for any risk as variable itself.
Do with gamma dist^{n} → $y \sim$ Poisson($\mu = \theta$), $\theta \sim$ gamma()
Results in y following -ve binomial dist^{n}, this restricts $\phi = 1$ but includes k – OD parameter.

-ve binomial variance f^{n} → $V(\mu) = \mu(1 + k\mu)$

k serves to 'inflate' variance over and above mean
As $k \to 0$, NB → Poisson

2.7.3 Tweedie distribution

To model premiums: most often 0, as most policies incur no losses; where losses, dist^{n} of losses highly skewed → done with Tweedie dist^{n}

As well as μ and ϕ, power parameter p - takes on all real numbers except those between 0 and 1

Variance f^{n} $V(\mu) = \mu^p$

$p=0$: normal. $p=1$: Poisson. $p=2$ Gamma

$p=3$: inverse Gaussian

Tweedie provides continuum of dist^{n}s between iG and G by setting p between 2, 3

We are interested in p between 1, 2 → ends of range are Poisson (frequency r) and Gamma (severity r) → ideal for modelling pure premium / loss ratio

Tweedie dist^{n} models a 'compound Poisson - Gamma process'. Events follow a Poisson process, each event generates a loss which follows a gamma dist^{n}.

Here λ = Poisson mean and variance, μ = Tweedie mean

2.7.3 Tweedie distribution

Gamma dist^{n} takes shape and scale parameters α and θ. The mean is ...

$$E[y] = \alpha \cdot \theta$$

Coefficient of variation is ... $CV = \frac{1}{\sqrt{\alpha}}$

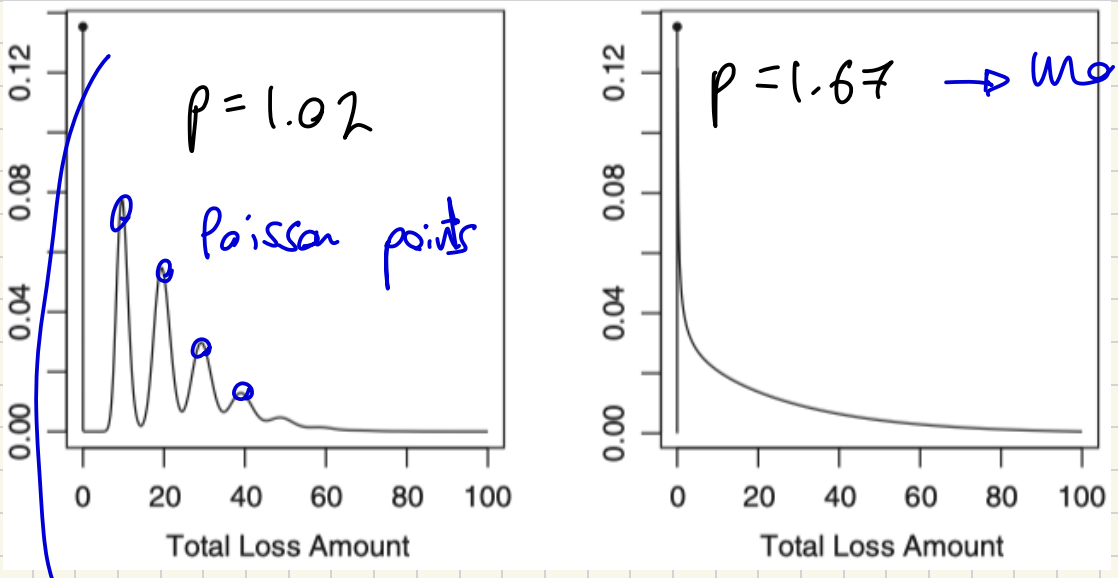
Tweedie mean related by = $E[y] = \mu = \lambda \cdot \alpha \cdot \theta$
(prod of Poisson mean and gamma mean)
(pure premium = freq × severity)

Power param $p = (\alpha + 2) / (\alpha + 1)$

↳ only f^{n} of gamma param α i.e. p (CV) exists

As $CV \to 0$, $p = 1$. $CV \to \infty$, $p = 0$

$1.5 \leq p \leq 1.8$ common for insurance → more realistic



No claims likely

2.7.3 Tweedie distribution

Tweedie dispersion parameter $\phi = \frac{\lambda^{1-p} \cdot (\alpha\theta)^{2-p}}{2-p}$

In Tweedie GLM, μ varies by record, controlled by linear predictor, while ϕ, p constant $\forall$ records. Implication → Tweedie GLM contains assumption that freq. and sev. move in same dirn.

Increase in target made up of increase in both freq. and sev. components.

How to determine p?

1. Model-fitting software
2. Test candidate values. Goal to optimise Log-likelihood
3. Select logical values → 1.6, 1.67, 1.7 (most practical)

2.8 Logistic regression

Sometimes target variable is dichotomous / binary – non numeric, but occurance of an event (will policy holder renew?)
Built on dataset of historical records of similar scenarios for which outcome is known. y_i = 1 (occur) or 2 (not occur) → Binomial dist^n

Linear predictor takes values $[-\infty, +\infty]$
Binomial dist^n must have $[0, 1]$
∴ Need link f^n to go $A \rightarrow B$
→ $g(\mu) = \ln[\mu / (1-\mu)]$ (Logit link f^n)

$\mu = 1/(1+e^{-x})$ (Logistic f^n – inverse)

translates linear predictor value onto probability
Large -ve LP = Low P(occur) vice versa

$y_i \sim bin(\mu_i)$

$\ln[\mu_i/(1-\mu_i)] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$

$odds = \frac{\mu}{1-\mu}$ → occurance / non-occurance — not bound by $[0, 1]$

↳ exp eq^n → multiplicative series of terms = odds of occurance

GLM coeff's describe effect of predictors on odds
Coeff = 0.24 → unit increase in x increases odds by $e^{0.24} - 1 = 27\%$

2.9 Correlation among predictors, multicollinearity and aliasing

Predictors can exhibit correlation among them - GLM good at finding unique effect of var.

Multicollinearity -> Correlation between pairs of predictors detectable with correlation matrix. Problem arises when 2+ predictors predictive of a third (multicol.).
Possible instability as variable may not be highly correlated with either var. individually.
Detect with VIF -> measure of how much squared standard error for predictor is increased due to collinearity. VIF > 10 is high

Aliasing -> two predictors perfectly correlated said to be aliased, GLM will not have unique sol^{n}.
Most GLM software will detect this and drop a predictor from model.
Nearly perfectly correlated, software may try run anyway. Extreme correlation = highly unstable -> failure to converge, nonsense coeff's.

2.10 Limitations of GLMs

① GLM assigns full credibility to the data

If we have a small dataset, GLM still fully trusts dataset rather than adjusting toward broader average.
Can adjust unstable estimates using credibility procedures or smoothing methods

② GLMs assume randomness of outcome is uncorrelated

When GLM predicts something, outcome split into two parts - part explained by model and random part.
GLM assumes one records random error does not affect another
Why not true? A risky driver one year will likely be a risky driver next year

∴ underestimates uncertainty / overstates confidence → coeff's or p-values too precise

Can fix with more advanced models...
GLMM and GEE