Generalized Linear Models

Lecture 12. Tweedie models. Compound Poisson models

Tweedie distributions $\text{Tw}(\mu, p, \varphi)$, $\text{Tw}_p(\mu, \varphi)$

Tweedie distributions form a subclass of exponential dispersion family, the class includes continuous distributions like normal, gamma and inverse gaussian, discrete distributions like Poisson, and also Poisson-gamma mixtures (that have positive probability at zero and are continuous elsewhere)

The variance function for Tweedie distributions has the following form $\nu(\mu) = \mu^p$

GLM with Tweedie response

Canonical parameter θ_i and the canonical function $b(\theta_i)$ are

$$\theta_i = \theta(\mu_i) = \begin{cases} \frac{\mu_i^{1-p}}{1-p}, & p \neq 1 \\ \log \mu_i, & p = 1 \end{cases} \qquad b(\theta_i) = \begin{cases} \frac{\mu(\theta_i)^{2-p}}{2-p}, & p \neq 2 \\ \log \mu(\theta_i), & p = 2 \end{cases}$$

- Mean: μ_i
- Variance: $\varphi\mu_i^p$ – *power-variance distributions*

The Tweedie distributions were named by Bent Jorgensen after Maurice Tweedie, a statistician and medical physicist at the University of Liverpool, UK, who presented the first thorough study of these distributions in 1984.

Tweedie family $\mathrm{Tw}_p(\mu, \varphi)$

- Tweedie distributions exist for all $p > 0$, no analytic form exists for $0 < p < 1$
- If $1 < p < 2$, the distribution are continuous for $Y > 0$, with a positive mass at $\mathrm{Y} = 0$.
- If $p > 2$, the distributions are continuous, $Y > 0$

Known distributions:

- $p = 0$ – normal distribution
- $p = 1$ and $\varphi = 1$ – Poisson distribution
- $1 < p < 2$ – compound Poisson-gamma distribution
- $p = 2$ – gamma distribution
- $p = 3$ – inverse gaussian distribution
- $2 < p < 3, p > 3$ – positive stable distributions

There are Tweedie models that allow for zero-issues as well

Applications of Tw_p distribution

1. $1 < p < 2$
   - Fish count estimation (silky shark, tuna fish): $\hat{p} = 1.12$, Shono (2008, 2010)
   - Root length density of apple trees: $\hat{p} = 1.4$, Silva (1999)
2. $p \geq 2$
   Tweedie distribution with $p \geq 2$ is a continuous non-negative distribution. It is quite similar to gamma distribution ($p = 2$), but is more skewed to the right (the bigger p, the more skewed)

   Examples:
   - Survival analysis of animals: $\hat{p} = 3.85$
   - Butterfat content in milk: $\hat{p} \approx 5$

Here p and $\hat{p}$ denote the Tweedie parameter and its estimate

Compound Poisson (CPo) models (1)

Compound Poisson distribution

Let $N \sim Po(\lambda)$, and let Z_i be some i.i.d. random variables independent of N. Then $Y = \sum_{i=1}^{N} Z_i$ has compound Poisson distribution.

A very commonly used CPo distribution is Poisson-gamma distribution, i.e. Tweedie distribution with $1 < p < 2$.

Tw_p $(1 < p < 2)$

Let us have $N \sim Po(\lambda)$ and $Z_i \sim \Gamma(\alpha, \gamma)$ are i.i.d., where γ is the scale parameter (inverse of rate, i.e. $\mathbf{E}Z_i = \alpha\gamma$). Then the distribution of $Y = \sum_{i=1}^{N} Z_i$ is Tweedie distribution with parameters:

$$p = \frac{\alpha + 2}{\alpha + 1}, \quad \mu = \lambda\alpha\gamma, \quad \varphi = \frac{\lambda^{1-p}(\alpha\gamma)^{2-p}}{2 - p}$$

and variance is $\varphi\mu^p = \lambda\gamma^2\alpha(\alpha + 1)$

Compound Poisson (CPo) models (2)

The parameters of Poisson and gamma distributions can be calculated from Tweedie parameters as follows:

- $\lambda = \frac{\mu^{2-p}}{\varphi(2-p)}$
- $\alpha = \frac{2-p}{p-1}$
- $\gamma = \varphi(p-1)\mu^{p-1}$

Most common link function used in Tweedie models is *log* (but it is not canonical link)

In R: library `tweedie` (also library `cplm`)

$p = 1.5$

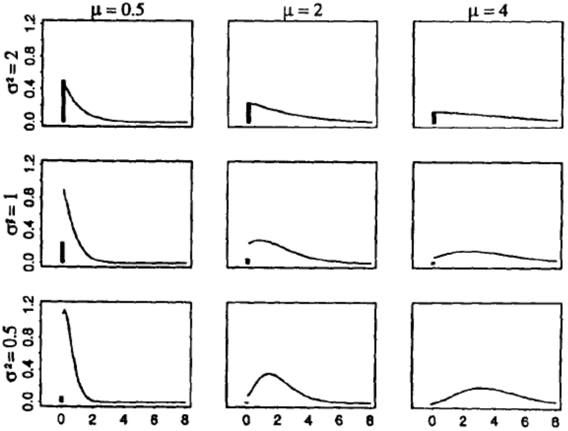


Figure 4.5. ***Some compound Poisson Tweedie density functions—note positive probability in zero***

Source: Kelly, Jorgensen (1996). Analyzing Accident Benefit Data Using Tweedie's Compound Poisson Model

Example. Non-life insurance claim payments (1)

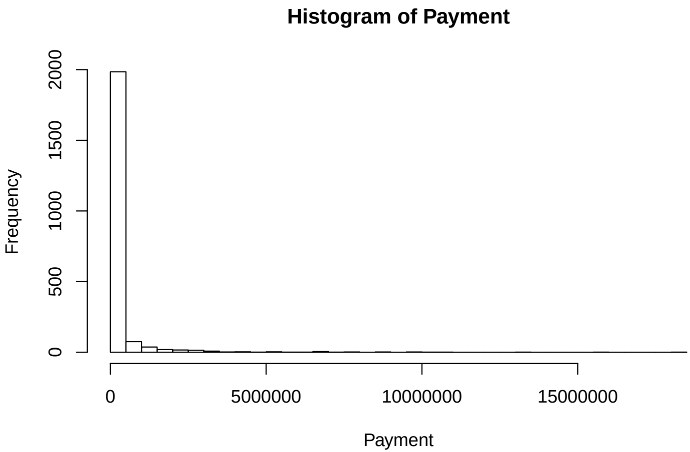
MTPL insurance claims in Sweden for the year 1977

In Sweden all motor insurance companies apply identical risk arguments to classify customers, and thus their portfolios and their claims statistics can be combined. The data were compiled by a Swedish Committee on the Analysis of Risk Premium in Motor Insurance. The Committee was asked to look into the problem of analyzing the real influence on claims of the risk arguments and to compare this structure with the actual tariff.

The dataset has 2182 observations and contains the following variables:

- Kilometres – Kilometres travelled per year (5 classes: 1: $<$ 1000, 2: 1000-15000, 3: 15000-20000, 4: 20000-25000, 5: $>$ 25000)
- Zone – Geographical zone (7 zones)
- Bonus – No claims bonus. Equal to the number of years, plus one, since last claim
- Make – 1-8 represent eight different common car models. All other models are combined in class 9
- Insured – Number of insured in policy-years
- Claims – Number of claims
- Payment – Total value of payments in SEK

Example. Non-life insurance claim payments (2)



Example. Solution in R (1)

```
> library(tweedie)
> formula = "Payment~as.factor(Kilometres)+as.factor(Zone)
             +as.factor(Bonus)+as.factor(Make)+offset(log(Insured))"
> p.vec = seq(1.2, 1.8, by=0.05)
> twp = tweedie.profile(formula, p.vec=p.vec, method="series",
                        do.plot=TRUE, verbose=TRUE, data = swautoins)
---
 This function may take some time to complete;
 Please be patient.  If it fails, try using  method="series"
 rather than the default  method="inversion"
 Another possible reason for failure is the range of p:
 Try a different input for  p.vec
---
1.2 1.25 1.3 1.35 1.4 1.45 1.5 1.55 1.6 1.65 1.7 1.75 1.8
p = 1.2
* Phi estimation, method:  mle (using optimize):  Done (phi = 1405.607 )
* Computing the log-likelihood (method = series ): Done: L = -21656.34
p = 1.25
* Phi estimation, method:  mle (using optimize):  Done (phi = 933.0532 )
* Computing the log-likelihood (method = series ): Done: L = -21519.04
...
```

Example. Solution in R (2)

```
> twp$p.max
[1] 1.359184
> library(statmod)
> model = glm(formula, family="tweedie"(var.power=twp$p.max,link.power=0),
              data=swautoins)
> summary(model)
...
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.60133    0.05339 123.646  < 2e-16 ***
as.factor(Kilometres)2  0.21347    0.03215   6.640 3.96e-11 ***
as.factor(Kilometres)3  0.31528    0.03504   8.997  < 2e-16 ***
as.factor(Kilometres)4  0.39158    0.04249   9.216  < 2e-16 ***
...
(Dispersion parameter for Tweedie family taken to be 558.0199)
    Null deviance: 1857340  on 2181  degrees of freedom
Residual deviance:  878626  on 2157  degrees of freedom
AIC: NA

> AICtweedie(model)
[1] 42924.11
```

Zero-inflated Compound Poisson model (ZICPo)

Excess zeros in insurance: no claims + small claims not reported

Claim counts modelled by ZIP, amounts modelled by gamma or inverse gaussian

If claim amounts follow gamma distribution, we have

Zero-inflated Poisson-gamma model

$$\begin{cases} 0, \text{ with probability } \pi \\ CPo(\mu, p, \varphi), \text{ with probability } 1-\pi \end{cases}$$

Zeros come from two sources (count process + zero process):

$$\mathbf{P}(Y=0) = \pi + (1-\pi)\exp\left(-\frac{\mu^{2-p}}{\varphi(2-p)}\right)$$

Idea is similar to ZIP model, but we have additional component (claim amount)

Zero-inflated model takes into account the '*bonus hunger*' problem: small claims may be not reported due to possible gain in premium next year.

Zero-adjusted Compound Poisson (ZACPo)

Claim count modelled by ZAP, individual claims modelled by gamma or inverse gaussian

Let π be probability of a claim, Z_i the claim amount

$$\begin{cases} 1-\pi, \text{ if } y = 0 \\ \pi g(z), \text{ if } y > 0 \end{cases}$$

where $g(z)$ is the pdf of claim amount
Resulting distribution has positive probability mass at 0!
Estimation:

- loss probability from *logit* model
- claim amount from *log* model

Example. Australian insurance claims 2004-2005

Australian insurance claims

Let us recall the car claims dataset used in Lecture slides 5. The dataset contains 10 variables and 67856 rows (insurance policies), from where 4624 policies actually had claims. The dataset contains the following variables

- veh_value – vehicle value
- exposure – risk exposure
- clm – claim indicator (0/1)
- numclaims – number of claims within the period
- claimcst0 – claim amount (0 when there is no claim)
- ...

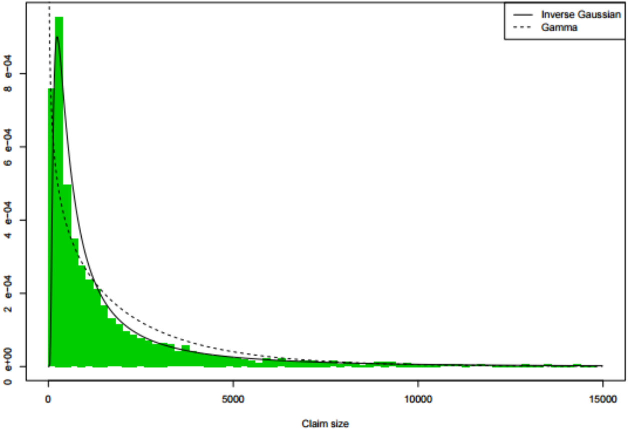
Candidate models for claim count were: Poisson, ZIP and negative binomial

Candidate models for claim amount were: gamma and inverse Gaussian

Interesting (or expected?) findings:

- The age of insured and the area affected both frequency and total payment
- For younger drivers, both frequency and payment were bigger

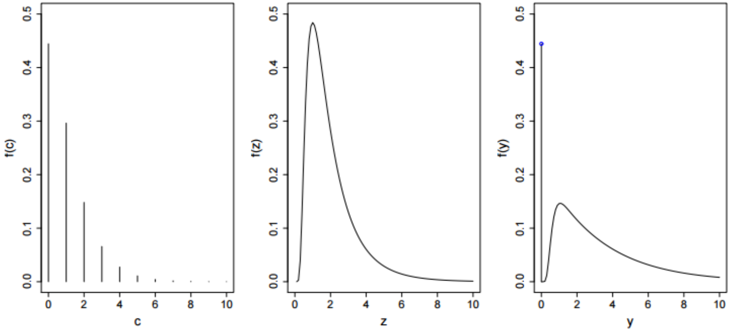
Example continued. Severity distribution



Australian insurance claims 2004-2005

Source: Heller et al. (2007) Mean and dispersion modelling for policy claims costs

Example continued. Insurance claim payments



Claim number (frequency) distribution: negative binomial, claim amount distribution: inverse Gaussian and claim payment distribution (cost per policy)

Source: Heller et al. (2007) Mean and dispersion modelling for policy claims costs

ZAIG model

Another alternative to Tweedie models is the zero-adjusted Inverse Gaussian (ZAIG) model

Consider the insurance framework and let Y_i be the size of claim of i-th policy

Then Y_i has a mixed discrete-continuous probability function:

$$f(y_i) = \begin{cases} 1 - \pi_i, & \text{if } y_i = 0, \\ \pi_i g(y_i), & \text{if } y_i > 0, \end{cases}$$

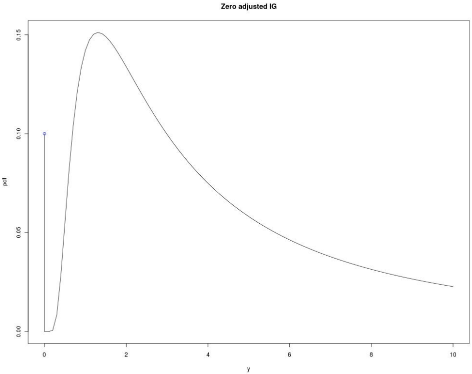
where

$$g(y_i; \mu_i, \lambda) = \exp\{-\frac{\lambda(y_i - \mu_i)^2}{2\mu_i^2 y_i} + \frac{1}{2}\ln(\frac{\lambda}{2\pi y_i^3})\}$$

The model consists of two parts:

- IG model for continuous part
- Discrete part is usually modelled by logistic regression (i.e. binary model with *logit* link

Example continued. ZAIG model



Distribution of claim payment with ZAIG model. Notice the probability mass at 0.

In R: library `gamlss`

Source: Heller et al. (2006) The zero-adjusted Inverse Gaussian distribution as a model for insurance claims

Applications of CPo distribution in climatology

Different applications of Tweedie model to weather data:

- Analysis of rainy days in Melbourne (1981-1990), $\hat{p} = 1.58$
- Analysis of snowfall data in Seattle (1906–1960), $\hat{p} = 1.52$
- Average wind speed in Ireland (1961–1978), $1 < \hat{p} < 2$

Here $\hat{p}$ is the estimated Tweedie index

Modelling rainfall data (Dunn et al, 1996, Lennox, 2003)

Model:

- Dry days (rainfall amount is 0) and rainy days (rainfall amount > 0, treated as continuous r.v.)
- Rainfall amount follows gamma distribution
- Rainy days follow Poisson distribution

Applications of CPo in sociology and medicine

Alcohol use among British teenagers (Gilchrist, Drinkwater (1999)

Cohort study among 16-17 year old teenagers, $n = 1545$
Model:

- number of events of consumption is Poisson distributed
- the amount drunk on each occasion is gamma distributed
- total amount is Tweedie distributed ($\hat{p} = 1.41$)

Time-Use Data, TUD-data (Dunn&Brown, 2011)

Longitudinal study of Australian children (march-nov 2004), 4-5 y.o. children, 3456 diaries with data
Analysis of time spent on different activities was conducted, data contains many zeros, the non-zero part is treated as continuous

- 3 response variables: time spent on watching TV, traveling and walking
- Arguments used: weekday, sex, number of children, income of family, education of parents, etc

Example. Time-use data

In average, 4-5 year old children spent

- more than 2 hours a day watching TV
- less than 1 hour traveling by car
- less than 15 minutes walking

Corresponding medians:

- TV: 1.8 hours
- car: 0.8 hours
- walking: 0 hours

Results of Tweedie model:

- Tweedie indices: tv: $\hat{p} = 1.18$, car: $\hat{p} = 1.19$, walking: $\hat{p} = 1.32$
- weekday, income, mother's job – significant for tv-use, not for others
- father's job not significant for walking
- child's sex not significant for traveling
- number of children significant for traveling

Tweedie model is compared by *tobit* model and linear model and considered better

Source: Brown, J.E., Dunn, P.K. (2011). Comparisons of Tobit, Linear, and Poisson-Gamma Regression Models. An Application of Time Use Data. *Sociological Methods and Research*, 40(3), 511–535

Summary. Compound Poisson models

Compound Poisson models:

- Compound Poisson distribution as a special case of Tweedie distribution ($1 < p < 2$), Poisson + Gamma
- General compound Poisson distribution (CPo): Poisson + (usually continuous) distribution
- ZICPo – zero-inflated compound Poisson distribution (in case of excess zeros)
- ZACPo – zero-altered compound Poisson distribution (in case of excess zeros, when zeros itself are not of interest)