Dispersion Estimates for Poisson and Tweedie Models
Lecture in Indonesia november 2014

Stig Rosenlund

I will give a review of the paper by me with this title from 2010 in *ASTIN Bulletin* **40** (1), 271–279. Please interrupt me with questions at any time!

The following are my main conclusions.

- The concept ODP = 'Overdispersed Poisson', applied to claim numbers, is unnecessary. It cannot mean anything else than either Mixed Poisson or Compound Poisson. To speak of 'Overdispersed Poisson' is to invite confusion as to whether you mean Mixed or Compound Poisson. These are two different things.

- Mixed Poisson: A variable X is Poisson(λ) conditional on a random variable λ. Then X is Mixed Poisson distributed.
- Compound Poisson: If $X = \sum_{i=1}^{N} Y_i$, with N Poisson and Y_i independent, identically distributed and independent of N, then X is Compound Poisson distributed. If Y_i is a positive integer-valued random variable, the distribution is called Generalized Poisson.

■ The χ^2-based Pearson ϕ-estimate (ϕ is defined next page) is unsuitable for GLM log link claim frequency analysis, provided we can identify claims in a Generalized Poisson process occurring at the same time as belonging together. If such identification is possible, we can count these claims as one and add their amounts to one amount. Then we retrieve the pure Poisson process with $\phi = 1$.

The model

In GLM log link theory for claim frequency, the ODP (Overdispersed Poisson) model is used. In this theory tariff cells u are combinations of categorical covariates, called arguments. Let N_u be the number of claims occurring in tariff cell u during some period of time. The mean and variance of N_u depend on an exposure e_u, namely

A. $\mathrm{E}[N_u] = \nu_u e_u$

B. $\mathrm{Var}[N_u] = \phi \nu_u e_u$

Here ν_u, called claim frequency, is multiplicative in the arguments. That is, ν_u is a product of a base constant and a factor for each argu-

ment. The number $\phi \geq 1$ is an unknown constant called the dispersion parameter. The same number applies for all u and for any time period regardless of length. This means that $\mathrm{Var}[N] = \phi \mathrm{E}[N]$ for any claim number N. For pure Poisson $\phi = 1$, while the case $\phi > 1$ is denoted overdispersion.

Three basic assumptions are made in this GLM theory, namely

1) Independence between insurance policies

2) Independence between disjoint time intervals (independent increments)

3) Exposure homogeneity

These assumptions imply the linear dependence of variance on exposure in **B** above. Without the independent increments property **B** is hard to justify. Time heterogeneity can be brought back to time homogeneity by the concept of operational time. It is just that the assumption **3**) is convenient for avoiding unnecessarily complicated notation.

The literature suggests the χ^2-based Pearson ϕ-estimate, which we denote $\hat{\phi}$. Let

n = number of tariff cells

r = n:o of free parameters = $1 + \sum_{\text{arguments}}$[(n:o of classes per argument) - 1]

$\hat{\nu}_u$ = estimate of the claim frequency ν_u in the GLM Poisson log link

model.

The number of degrees of freedom is $n - r$. It holds

$$\hat{\phi} = (n-r)^{-1} \sum_{u=1}^{n} e_u \left(\frac{N_u}{e_u} - \hat{\nu}_u \right)^2 / \hat{\nu}_u$$

Now that you have seen the formula, **please forget it**! An insurance company that has a minimal check on its claims will not need it, as I have explained. Just set $\phi = 1$.

I showed in my paper that the Overdispersed Poisson model, as stated above, is the same as Compound Poisson. The Generalized Poisson model is a special case of Compound Poisson. But on Wikipedia I

have seen the definition of Overdispersed Poisson as being the same as Mixed Poisson – probably written by someone who was as confused as I was, when I started to investigate the matter.

Now for Compound Poisson we have $\phi > 1$, but in tariff analysis it is not interesting to state the problem in terms of the parameter ϕ. Besides, you can estimate it better than with the Pearson ϕ-estimate.

Why $\hat{\phi}$ and why the concept Overdispersed Poisson?

I think we can attribute it to

McCullagh, P. and Nelder, J. A. (1989), *Generalized linear models, Second Edition*,

and to

Renshaw, A. E. (1994), Modelling the claims process in the presence of covariates, *ASTIN Bulletin* **24(2)**, 265–285.

Renshaw suggested this mechanism to generate Overdispersed Poisson: Claims are generated by processes that are Poisson, conditional on random independent claim frequencies λ_u. In **A** and **B** above we would then have $\nu_u = \mathrm{E}[\lambda_u]$. But I could show that Renshaw's calculations were ambiguious. (I originally wrote that they were in errror, but a referee thought that was too critical.) The ambiguity gives rise to the apparent paradox that random claim arrival rates generate ODP processes with $\phi > 1$ having the independent increments property, while time-

homogeneous unit-step jump processes with independent increments are pure Poisson. This you can read in elementary text books on stochastic processes. In straightening out the ambiguity I could show that the asymptotic theory for confidence intervals in the GLM ODP log link theory cannot be applied to the random intensities case. This theory presupposes that $\mathrm{Var}[N_u/e_u] = \phi\nu_u/e_u \to 0$ as $e_u \to \infty$. But this is not so with random intensities.

I have shown the Overdispersed Poisson concept to be at best superfluous and at worst confusing. Still you can see it used in articles, as if the authors have not read or understood my article.

Why the Poisson process as model for claim numbers?

In claim frequency analysis of mass consumer insurance one can apply a general limit theorem for superpositions (sums) of point processes by

Grigelionis, B. (1963), On the convergence of sums of random step processes to a Poisson process, *Probability Theory and its Applications*, **8(2)**, 177–182.

The theorem states that under weak conditions the superposition of many independent unit-step claim occurrence processes, each one contributing a small part to the total, is approximately Poisson. This holds even for random intensities. For instance, when analyzing a portfolio of 60,000 customers with variances of the same order of magnitude, the in-

troduction of 60,000 random independent intensities for conditional Poisson processes is an unnecessary complication. For practical purposes, the pure Poisson assumption will give the same results. You can also read Appendix 8 in Rappmane.doc on the Rapp site. This essay cannot be published, since I have no new results.

Macroscopic fluctuations

Observed claim frequencies are often found to fluctuate more from year to year than what follows from the Poisson assumption. This holds also for mass consumer insurance. This is due to macroscopic variables (e. g. crime waves, business cycles, the weather) affecting large parts of the

portfolio in the same way. Here the assumption **1**) of independence between policies does not hold. So, for analyzing collective claim frequencies in mass consumer insurance, the model of random independent claim frequencies gives no help.

For analyzing price relativities, my 30-year experience with practical pricing is that it is mostly best to condition with respect to these macroscopic variables.

So we retrieve the Poisson process (although time-heterogeneous). It is seldom feasible to model how the effects of the macroscopics differ between tariff cells. Relying on e. g. theft expert judgments is better than augmenting the mathematical model.

New dispersion parameter estimate in the Tweedie model

My article concludes with a new dispersion parameter estimate in the GLM Tweedie model for risk premium, and a comparison by simulations between it and the Pearson estimate. When the exponent $p > 1$ in the Tweedie model, my new estimate is better, if there are sufficiently many claims in each tariff cell. Otherwise the Pearson estimate is better. For $p = 1$ – the Compound Poisson model – my estimate is always better. I am not rendering the formula here, because

1. It is not interesting to do tariff analysis in terms of the parameter ϕ.

2. The Tweedie model for risk premium should not be used.

The second reason is shown in my article

Rosenlund, S. (2014), Inference in multiplicative pricing, *Scandinavian Actuarial Journal* **2014**(8), 690-713.
http://www.tandfonline.com/doi/abs/10.1080/03461238.2012.760885.