DISPERSION ESTIMATES FOR POISSON AND TWEEDIE MODELS

BY

STIG ROSENLUND

ABSTRACT

As a consequence of pointing out an ambiguity in Renshaw (1994), we show that the Overdispersed Poisson model cannot be generated by random independent intensities. Hence Pearson's chi-square-based estimate is normally unsuitable for GLM (Generalized Linear Model) log link claim frequency analysis in insurance. We propose a new dispersion parameter estimate in the GLM Tweedie model for risk premium. This is better than the Pearson estimate, if there are sufficiently many claims in each tariff cell. Simulation results are given showing the differences between it and the Pearson estimate.

KEYWORDS

Generalized Linear Model, GLM log link, ODP, Overdispersed Poisson, Tweedie.

1. MODEL AND PERSPECTIVES FOR CLAIM FREQUENCY

1.1. 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. $\mathrm{E}[N_u] = v_u e_u$
B. $\mathrm{Var}[N_u] = \phi v_u e_u$

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

Astin Bulletin 40(1), 271-279. doi: 10.2143/AST.40.1.2049229

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

See e.g. Ohlsson & Johansson (2010), section 1.2. 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.

In section (6.2.4) of McCullagh & Nelder (1989) the χ^2-based Pearson ϕ-estimate is suggested. Renshaw (1994) applies GLMs to multiplicative models in insurance. The Pearson estimate (2.16) of ϕ is there denoted $\hat{\phi}$. Let

n = number of tariff cells
r = n:o of free parameters = $1 + \Sigma_{\text{arguments}}$ [(n:o of classes per argument) – 1]
$\hat{v}_u$ = estimate of the claim frequency v_u in the GLM Poisson log link model.

The number of degrees of freedom is $n - r$, denoted v in Renshaw (1994). It holds

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

1.2. Generalized Poisson

Consider the Generalized Poisson case $N_u = \sum_{i=1}^{A_u} Z_{ui}$. Here A_u is Poisson and independent of Z_{ui}. The Z_{ui} are, for a specific u, IID positive integer random variables. They count claims occurring at the same time from the same cause. This is an ODP model, provided that $\phi = \mathrm{E}[Z_{u1}^2]/\mathrm{E}[Z_{u1}]$ is the same for different u. Assume that the Zs can be observed directly. Then it will follow from sections 3 and 4, by specializing the Tweedie model to $p = 1$, that the simple quotient $\sum_{u,i} Z_{ui}^2 / \sum_{u,i} Z_{ui}$ is preferable to the Pearson estimate $\hat{\phi}$. If direct observation is not possible, and if we wrongly assume that the Z_{ui} claims occurring at the same time arrive in an ordinary Poisson process, then it is wrong to use the dispersion parameter $\phi = 1$. The Pearson estimate is then useful.

Cases in insurance where the Zs cannot be observed directly are rare. For example regarding claims from storm damage, great care is traditionally taken to ascertain direct observability by identifying simultaneous claims arising from the same cause (storm). The practical actuarial handling in insurance, for the purpose of variance estimates, is to add the claim amounts X_{uij} associated with Z_{ui} to a sum $X_{ui} = \sum_{j=1}^{Z_{ui}} X_{uij}$ per time point. Then these simultaneous claims count as a single claim. So we retrieve the pure Poisson process with $\phi = 1$ for claim occurrences.

https://doi.org/10.2143/AST.40.1.2049229 Published online by Cambridge University Press

1.3. Random independent claim frequencies

Another mechanism to generate ODP is suggested in Renshaw (1994), section 3. Namely that claims are generated by processes that are Poisson, conditional on random independent claim frequencies λ_u. In **A** and **B** above then $v_u = \mathrm{E}[\lambda_u]$. However, that the ODP model seems to follow from Renshaw's calculations is due to an ambiguity, described in the next section. 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, see Parzen (1962), 4-2. In straightening out the ambiguity we can see 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 $\mathrm{Var}[N_u / e_u] = \phi v_u / e_u \to 0$ as $e_u \to \infty$. But this is not so with random intensities, see (2) below.

1.4. Superpositions of many independent point processes

On the other hand, in collective claim frequency analysis of mass consumer insurance one can apply a general limit theorem for superpositions (sums) of point processes by Grigelionis (1963). This 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 introduction 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.

1.5. Random intensities in bonus/malus analysis

The analysis of individual customer claim frequencies for bonus/malus purposes is another matter. There the model of random intensities, Γ-distributed for convenience, is useful.

1.6. 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, our 25-year experience with practical pricing is that it is mostly best to condition with respect to these macroscopic variables.

https://doi.org/10.2143/AST.40.1.2049229 Published online by Cambridge University Press

Thereby 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.

2. Renshaw's ambiguity on random intensities

Independent response variables Y_u are defined in Renshaw (1994) for tariff cells u. For claim frequency analysis claim numbers N_u, exposures e_u and (possibly stochastic) claim rates λ_u are introduced. On p. 271 line 8 in Renshaw (1994) the responses are defined as $Y_u = N_u$. On p. 272 line 23 the notation is changed to $Y_u = N_u/e_u$. Renshaw writes "Focus on the weighted Poisson responses $Y_u (= N_u/e_u)$ with $Y_u \sim \text{Poi}(\lambda_u)$ so that

$$\text{E}(Y_u) = \text{E}\{\text{E}(Y_u|\lambda_u)\} = \text{E}(\lambda_u), \tag{3.2}$$

$$\text{Var}(Y_u) = \text{E}\{\text{Var}(Y_u|\lambda_u)\} + \text{Var}\{\text{E}(Y_u|\lambda_u)\}$$

and hence

$$\text{Var}(Y_u) = \text{E}(\lambda_u) + \text{Var}(\lambda_u)." \tag{3.3}$$

Both parts of (3.2) are correct. Assuming $Y_u|\lambda_u \sim \text{Poi}(\lambda_u)$, then (3.3) is also correct. But this assumption is not, unless $e_u = 1$, consistent with the definition $Y_u = N_u/e_u$ and not with the word "weighted". Because on p. 271, lines 4-7, N_u was defined as a random claim number, with realization n_u, such that $N_u|\lambda_u \sim \text{Poi}(e_u\lambda_u)$. Hence there is an ambiguity as to what Y_u is. The definition given for N_u and the subsequent definition $Y_u = N_u/e_u$ are necessary for an investigation using (3.2) of whether the ODP model holds for random intensities. Adhering to these definitions, we will show that (3.3) must be corrected. A crucial factor $1/e_u$ is missing in the first term of the right side of (3.3). A corrected version of (3.3) is as follows.

The first term of the right side of the second part of (3.2):

$$\text{Var}[N_u|\lambda_u] = e_u\lambda_u$$

$$\text{Var}[Y_u|\lambda_u] = \text{Var}[N_u/e_u|\lambda_u] = e_u\lambda_u/e_u^2 = \lambda_u/e_u$$

$$\text{E}[\text{Var}[Y_u|\lambda_u]] = \text{E}[\lambda_u]/e_u$$

The second term of the right side of the second part of (3.2):

$$\text{E}[Y_u|\lambda_u] = \text{E}[N_u/e_u|\lambda_u] = e_u\lambda_u/e_u = \lambda_u$$

$$\text{Var}[\text{E}[Y_u|\lambda_u]] = \text{Var}[\lambda_u] \text{ (as correctly given in (3.3))}$$

https://doi.org/10.2143/AST.40.1.2049229 Published online by Cambridge University Press

and hence

$$\mathrm{Var}[Y_u] = \mathrm{Var}[N_u/e_u] = \mathrm{E}[\lambda_u]/e_u + \mathrm{Var}[\lambda_u] \tag{2}$$

$$\mathrm{Var}[N_u] = e_u^2\,\mathrm{Var}[Y_u] = e_u\mathrm{E}[\lambda_u] + e_u^2\,\mathrm{Var}[\lambda_u]$$

$$\mathrm{Var}[N_u]/\mathrm{E}[N_u] = 1 + e_u\mathrm{Var}[\lambda_u]/\mathrm{E}[\lambda_u] \tag{3}$$

Expression (2) does not $\to 0$ as $e_u \to \infty$, unless $\mathrm{Var}[\lambda_u] = 0$.

From the correction just made it follows that random intensities, while entailing $\mathrm{Var}[N] > \mathrm{E}[N]$ for any claim number N, does not give the ODP model, since this model assumes that the left side of eq. (3) is a constant ϕ, the same for all u. If the λ_u are IID, the expression (3) would be larger for larger exposures e_u. Renshaw's expression (3.3) together with mistaking Y_u for N_u (the first definition of Y_u), on the other hand, implies the same constant ϕ for all u in the left side of eq. (3).

3. New dispersion parameter estimate in Tweedie's risk premium model

3.1. Model

In the Tweedie model for risk premiums with exponent p, the assumptions **1**), **2**) and **3**) of section 1.1 are supposed to be true. In addition to the definitions of section 1.1, let

$X_{ui}\,(i = 1, \ldots, N_u)$ = independent claim amounts, distributed as X_{u1} in class u

$S_u = \sum_{i=1}^{N_u} X_{ui}$

$\tau_u = \mathrm{E}[S_u/e_u]$ = risk premium

$\hat{\tau}_u$ = estimate of the risk premium τ_u in the GLM Tweedie log link model

The model is the following. For $\phi \geq 1$, the same for all u and for any time period,

$$\mathrm{Var}[S_u] = \phi e_u \mathrm{E}[S_u/e_u]^p \quad \text{or equivalently} \quad \mathrm{Var}[S_u/e_u] = \phi \mathrm{E}[S_u/e_u]^p/e_u \tag{4}$$

See Jørgensen & Paes de Souza (1994). As pointed out by Venter (2007), section 4.1, the link between claim frequency and claim severity is problematic in this model. The Pearson estimate is

$$\hat{\phi} = (n-r)^{-1} \sum_{u=1}^{n} e_u \left(\frac{S_u}{e_u} - \hat{\tau}_u \right)^2 / \hat{\tau}_u^p \tag{5}$$

https://doi.org/10.2143/AST.40.1.2049229 Published online by Cambridge University Press

3.2. New dispersion parameter estimate

If the claim occurrence processes are unit-step (one claim at a time) they are pure Poisson, as follows from the preceding sections. So, assuming unit-step, $\hat{\phi}$ does not have an advantage by catching a possible overdispersion in the claim occurrence processes. A ϕ-estimate utilizing that the claim occurrence processes are Poisson is useful, i.e. sometimes better than $\hat{\phi}$. (An appropriate measure of goodness is the mean square deviation of the estimate from ϕ.) We propose such an estimate. It will be better if there are sufficiently many claims in all tariff cells that have claims.

From (4) we get

$$\phi = e_u^{p-1}\,\mathrm{E}[S_u]^{-p}\,\mathrm{Var}[S_u]$$

The mean and variance of these Compound Poisson distributions are

$$\mathrm{E}[S_u] = \mathrm{E}[N_u]\,\mathrm{E}[X_{u1}] \quad \mathrm{Var}[S_u] = \mathrm{E}[N_u]\,\mathrm{E}[X_{u1}^2]$$

Hence for any u this is the (problematic) link between frequency and severity:

$$\phi = e_u^{p-1}\,\mathrm{E}[N_u]^{-p}\,\mathrm{E}[X_{u1}]^{-p}\,\mathrm{E}[N_u]\,\mathrm{E}[X_{u1}^2] = \mathrm{E}[N_u/e_u]^{1-p}\,\mathrm{E}[X_{u1}]^{-p}\,\mathrm{E}[X_{u1}^2]$$

This suggests a u-specific ϕ-estimate

$$\hat{\phi}_u = (N_u/e_u)^{1-p}\left(\frac{1}{N_u}\sum_{i=1}^{N_u} X_{ui}\right)^{-p}\left(\frac{1}{N_u}\sum_{i=1}^{N_u} X_{ui}^2\right) = \left[e_u\left(S_u/e_u\right)^p\right]^{-1}\sum_{i=1}^{N_u} X_{ui}^2 \qquad (6)$$

which will converge a. s. to ϕ when $e_u \to \infty$ as time $\to \infty$. The $\hat{\phi}_u$ are independent, so a linear combination $\sum_{u=1}^{n}\alpha_u\hat{\phi}_u$ with $\alpha_u \geq 0$ and $\sum_{u=1}^{n}\alpha_u = 1$ can give a better estimate than any single $\hat{\phi}_u$. The standard solution for this situation, which gives the estimate the smallest variance, is $\alpha_u \propto 1/\mathrm{Var}[\hat{\phi}_u]$. Here $\mathrm{Var}[\hat{\phi}_u]$ must be estimated, which is difficult to do exactly. We have attempted approximations for large $\mathrm{E}[N_u]$. The resulting ϕ-estimate was only marginally better than the estimate below in expression (8), when all expected numbers of claims $\mathrm{E}[N_u] = v_u e_u$ per tariff cell were large (the limiting case $e_u \to \infty$). And when $v_u e_u$ were small, then it was worse, even for $p = 1$.

We propose the following weights, appropriately larger for cells with more claims,

$$\alpha_u = e_u\left(S_u/e_u\right)^p\left(\sum_{j=1}^{n} e_j\left(S_j/e_j\right)^p\right)^{-1} \qquad (7)$$

so that our proposed new ϕ-estimate, converging almost surely to ϕ when $e_u \to \infty$, is

https://doi.org/10.2143/AST.40.1.2049229 Published online by Cambridge University Press

$$\hat{\phi}_0 = \sum_{u=1}^{n} \alpha_u \hat{\phi}_u = \frac{\sum_{u=1}^{n} \sum_{i=1}^{N_u} X_{ui}^2}{\sum_{u=1}^{n} e_u^{1-p} S_u^p} \tag{8}$$

3.3. Dispersion parameter estimate for Overdispersed Poisson

For the Overdispersed Poisson model $p = 1$ the expression (8) specializes to

$$\hat{\phi}_0 = \frac{\sum_{u=1}^{n} \sum_{i=1}^{N_u} X_{ui}^2}{\sum_{u=1}^{n} \sum_{i=1}^{N_u} X_{ui}} \quad (p = 1) \tag{9}$$

Here $\hat{\phi}_0 = U_2/U_1$ with $U_1 = \sum_{u=1}^{n} \sum_{i=1}^{N_u} X_{ui}$ and $U_2 = \sum_{u=1}^{n} \sum_{i=1}^{N_u} X_{ui}^2$. If the total number of claims $\sum_{u=1}^{n} N_u$ is moderately large, then with high probability $U_1 \approx \mathrm{E}[U_1]$ and $U_2 \approx \mathrm{E}[U_2] = \phi \mathrm{E}[U_1]$. Hence $\hat{\phi}_0$ is a good estimate for $p = 1$, even if every tariff cell has at most one claim. That is not true for $p > 1$, as is shown by the simulation result Case 4 of the next section. If we let all $X_{ui} = 1$ in addition to $p = 1$, we get $\mathrm{Var}[N_u] = \phi \mathrm{E}[N_u]$ as in section 1.1. Then $\hat{\phi}_0$ simplifies to 1, as it should.

4. Comparison of dispersion parameter estimates in the Tweedie model

4.1. Remarks on estimate properties

First a few remarks on the differences between the χ^2 Pearson type estimate $\hat{\phi}$ and our new estimate $\hat{\phi}_0$. We consider $\hat{\phi}$ as written in expression (5). Generalizations are possible by subdivision of the time interval covered by the analysis in several intervals and/or by taking individual policy periods as the summands in (5). However, a practical consideration against that is that seasonal variations in claim frequency will enlarge the estimate undesirably. Policies will have renewal dates scattered over the year and will often have less than yearlong time periods that would be parts of a generalized $\hat{\phi}$, particularly if calendar year is an argument. Retrieving time homogeneity from heterogeneity by operational time might not be feasible.

We can then list these differences which argue for our new estimate.

(i) $\hat{\phi}$ is not defined for only one argument. In contrast $\hat{\phi}_0$ is.

(ii) $\hat{\phi}$ uses only the aggregated data S_u, leading to unnecessarily large variance of $\hat{\phi}$. This is analogous to using only the between-cell variation and throw away the within-cell variation in ANOVA (analysis of variance). In contrast $\hat{\phi}_0$ uses also the sums of squares X_{ui}^2.

https://doi.org/10.2143/AST.40.1.2049229 Published online by Cambridge University Press

On the other hand, for $p > 1$ the following point argues for the Pearson estimate $\hat{\phi}$.

(iii) $\hat{\phi}$ works well when there are so many tariff cells that each one has only a few claims, while $\hat{\phi}_0$ does not for $p > 1$. With more than, say, 15 arguments, there will typically be at most one claim in a tariff cell. The almost sure convergence when $e_u \to \infty$ as time $\to \infty$ described above does not apply to this situation.

But then again, when each tariff cell has at most one claim and at most one (partial) policy period, seasonal variations will enlarge $\hat{\phi}$ undesirably.

4.2. Simulation results

Secondly the results of a simulation study. We generated independent Poisson claim numbers N_u and independent Γ-distributed claims X_{ui} with mean $\alpha\theta_u$ and variance $\alpha\theta_u^2$. For all cases we set $\alpha = 1$. Multiplicative claim frequencies and mean claims obeying the Tweedie assumptions were used. For each of six cases, we give here results of five simulated samples and subsequent observations of the two estimates (pairwise on the same sample). GLM Tweedie log link equation solutions were made for each sample, since the Pearson estimate requires this, by expression (5). The observations are given as percentages of the true value. The latter is unimportant in itself in this context. The five observation pairs give interesting information on the pros and cons of the two estimates.

$p = 1$ (ODP)

CASE 1. $n = 4\,826\,809$. $\sum_{u=1}^{n} N_u \approx 2\,400\,000$. Typical $N_u = 1$, if > 0.

$\hat{\phi}$ and $\hat{\phi}_0$ have mean ϕ and almost the same small variance.

CASE 2. $n = 216$. $\sum_{u=1}^{n} N_u \approx 20\,000$. Typical $N_u \approx 100$.

$100\hat{\phi}/\phi$	103	106	103	97	105
$100\hat{\phi}_0/\phi$	99	100	100	100	100

CASE 3. $n = 27$. $\sum_{u=1}^{n} N_u \approx 8\,000$. Typical $N_u \approx 300$.

$100\hat{\phi}/\phi$	145	138	205	182	113
$100\hat{\phi}_0/\phi$	100	101	100	101	100

$p = 1.5$ (Tweedie)

CASE 4. $n = 4\,826\,809$. $\sum_{u=1}^{n} N_u \approx 550\,000$. Typical $N_u = 1$, if > 0.

$100\hat{\phi}/\phi$	100	100	100	101	100
$100\hat{\phi}_0/\phi$	52	51	53	51	51

https://doi.org/10.2143/AST.40.1.2049229 Published online by Cambridge University Press

CASE 5. $n = 216$. $\sum_{u=1}^{n} N_u \approx 74\,000$. Typical $N_u \approx 350$.

$100\hat{\phi}/\phi$	84	93	98	114	114
$100\hat{\phi}_0/\phi$	100	100	100	99	100

CASE 6. $n = 27$. $\sum_{u=1}^{n} N_u \approx 8\,000$. Typical $N_u \approx 300$.

$100\hat{\phi}/\phi$	120	132	112	126	36
$100\hat{\phi}_0/\phi$	102	101	102	101	97

ACKNOWLEDGEMENTS

Thanks are due to two referees for many valuable suggestions.

REFERENCES

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.

JØRGENSEN, B. and PAES DE SOUZA, M.C. (1994) Fitting Tweedie's Compound Poisson Model to insurance claim data. *Scandinavian Actuarial Journal*, **1994(1)**, 69-93.

MCCULLAGH, P. and NELDER, J.A. (1989) *Generalized linear models, Second Edition*. Chapman and Hall, Boca Raton.

OHLSSON, E. and JOHANSSON, B. (2010) *Non-Life Insurance Pricing with Generalized Linear Models*. Springer.

PARZEN, E. (1962) *Stochastic Processes*. Holden-Day, San Francisco.

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

VENTER, G.G. (2007) Generalized Linear Models beyond the Exponential Family with Loss Reserve Applications. *ASTIN Bulletin* **37(2)**, 345-364.

STIG ROSENLUND
Västmannagatan 93
S-113 43 Stockholm
Sweden
E-Mail: stig.rosenlund@sverige.nu