# Pure Premium Regression with the Tweedie Model

Fitting regression models to insurance loss data has always been problematic. The problem is particularly acute for data from individual insurance policies where most of the losses are zero, and for those policies with a positive loss, the losses are highly skewed. Most of the traditional regression models do not deal with a mixture of discrete losses of zero and continuous positive losses. One way of dealing with this problem is to fit separate models to the frequency and severity, and estimate the pure premium by multiplying the result of each model. One can take issue with assumption of "separate" models.

Gordon Smyth and Bent Jørgensen provide an interesting alternative in their paper "Fitting Tweedie's Compound Poisson Model to Insurance Claims Data," which appeared in the 2002 *ASTIN Bulletin*. They first characterize the Tweedie model as a compound Poisson distribution. This distribution can be viewed as a computer simulation.

1. Select a random claim count, $N$, from a Poisson distribution with mean $\lambda$.
2. If $N > 0$, for $i = 1,..,N$, select a random claim size, $Z_i$, from a gamma distribution with scale parameter, $\theta > 0$ and shape parameter, $\alpha > 0$
3. If $N > 0$, set the loss, $X = \Sigma Z_i$. If $N = 0$, set $X = 0$.

The usual parameterization of the Tweedie distribution is given by $\mu$, $\phi$ and $p$, where the expected value of $X$ is equal to $\mu$ and the variance of $X$ is equal to $\phi \cdot \mu^p$. Smyth and Jørgensen translate the parameters of the compound Poisson distribution into the usual Tweedie parameters as follows:

$$\mu = \lambda \cdot \alpha \cdot \theta, \quad p = \frac{\alpha + 2}{\alpha + 1} \text{ and } \phi = \frac{\lambda^{1-p} \cdot (\alpha \cdot \theta)^{2-p}}{2 - p}. \qquad (1)$$

One can see from the middle equation that $p$ will be between 1 and 2. It is interesting to note that $p$ depends only on the shape parameter, $\alpha$, of the claim severity distribution. Also, since the coefficient of variation for a gamma distribution is equal to $1/\sqrt{\alpha}$, a claim severity distribution with losses clustered close to its mean value will have a high value of $\alpha$, and hence $p$ should be close to one. In my experience I typically find that the coefficient of variation for claim severity is greater than one, so we should expect $p$ to be greater than 1.5.

Figure 1 illustrates the connection between the compound Poisson and the Tweedie distributions. It shows a histogram of a 10,000 simulated losses and the density function of the corresponding Tweedie with a typical $\alpha$. Figure 2 shows the results of a similar exercise with a large $\alpha$. Here we see that the coefficient of variation for the claim severity distribution is small and the losses cluster around integral multiples of $\phi$ corresponding to claim counts of 0, 1, 2, … If we let $\alpha$ continue to increase indefinitely, $p$ approaches 1 and we get an overdispersed Poisson distribution where the only loss amounts with positive probability are integral multiples of $\phi$.

Let's now consider the Tweedie distribution in the context of regression modeling. As an illustration, I constructed a simple example with 50,000 observations where the frequency and severity means depend upon two independent variables, $x_1$ and $x_2$. In constructing this example, I was thinking of my real-world experience with auto insurance where $\lambda$ is small (around 0.05) and thus most policies have no losses. With the intent of fitting a GLM model with a log link, I simulated the losses from a compound Poisson distribution with the following parameters.

$$\begin{aligned} \log(\lambda) &= \log(0.05) + x_1 + 0.25 \cdot x_2, \\ \log(\theta) &= \log(10) + 0.25 \cdot x_1 + x_2, \text{ and } \alpha = 0.5 \end{aligned} \qquad (2)$$

Putting these equations together we have:

$$\log(\mu) = \log(\alpha) + \log(\lambda) + \log(\theta) = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 \qquad (3)$$

with $$\begin{aligned} a_0 &= \log(0.5) + \log(0.05) + \log(10) = -1.386, \\ a_1 &= 1.25, \text{ and } a_2 = 1.25. \end{aligned}$$

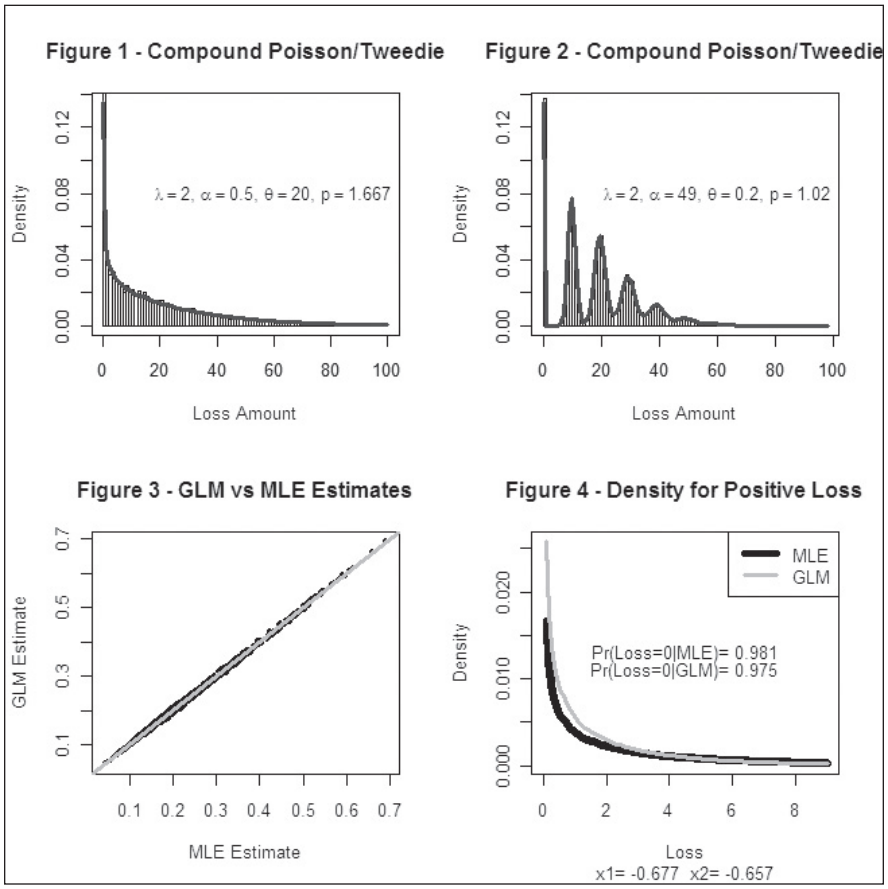
I then proceeded to fit some models to the simulated data observing only the simulated losses and corresponding values of $x_1$ and $x_2$.

First I estimated the $p$ parameter for the Tweedie model. Since $p$ depends only on the shape parameter of the gamma distribution, it can be estimated by fitting a GLM with a gamma distribution and a log link to the positive losses as follows[1].

$$Var[Loss] = \frac{\mu^2}{\alpha} = \phi \cdot \mu^2 \Rightarrow \alpha = 1/\phi \Rightarrow p = \frac{2 + 1/\phi}{1 + 1/\phi} \qquad (4)$$

The GLM fit to the simulated data with positive losses gave the value, $\phi = 2.135$. Equation 4 then gave an estimate of $p = 1.681$, which is close to the underlying model's parameter value of 1.667.

[1] To be technically correct, one should fit the model to individual claims. But since only 2% of the losses involved multiple claims it is a good approximation to fit the model total losses.




Next I fit a model of the form $\log(\mu) = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2$ with a Tweedie GLM and obtained the following results.

| Coefficients | Estimate | Std. Error Error | *t* value | Pr(>\|*t*\|) |
|---|---|---|---|---|
| Intercept | -1.41908 | 0.03786 | -37.486 | <2*e*-16 |
| $x_1$ | 1.21028 | 0.12639 | 9.576 | <2*e*-16 |
| $x_2$ | 1.39392 | 0.12705 | 10.972 | <2*e*-16 |

Note that these estimates are well within two standard errors of the corresponding parameter values in the underlying model.

A problem with using a GLM to fit a compound Poisson model is that the GLM assumes that the dispersion parameter, $\phi$, is constant. But if we believe our data has a compound Poisson distribution, according to Equation (1), $\phi$ varies with $\lambda$ and $\theta$. Smyth and Jørgensen devote a fair amount of efforts to addressing this by using a "Double GLM." I elected to take the conceptually simpler, but more computationally intense, approach of brute force maximum likelihood.[2] I fit a model in the form of Equation (2), fitting the equations for $\lambda$ and $\theta$ simultaneously, with the following results.

[2] For more information on this see my August 2008 "Brainstorms" column in the *Actuarial Review*.

| Coefficients | log($\lambda$) | log($\theta$) | Combined (with log($\alpha$)) |
|---|---|---|---|
| Intercept | -3.03364 | 2.38269 | -1.40954 |
| $x_1$ | 0.34877 | 0.79284 | 1.14161 |
| $x_2$ | 1.04565 | 0.38659 | 1.43224 |

To compare the results of the two models, I plotted the predictions of $\mu$ for a sample of 500 points in Figure 3, which shows close agreement between the two models. In Figure 4, I selected a random observation and plotted the interesting part of the Tweedie density functions implied by the $\mu$, $p$ and $\phi$ parameters. The difference in densities is small, but noticeable.

Was I lucky to get such close agreement? I think not. Heuristically, my reasoning is that the $\phi$ parameter is not in the expression for $\mu$, but it is in the expression for the variance, $\phi \cdot \mu^p$. Getting the variance right increases the efficiency of the estimator but has no direct effect on the estimate itself. When we have a large number of observations, in practice often running into the millions, efficiency is not terribly important. So for most applications I feel comfortable using estimates produced using a constant dispersion parameter that can be estimated with GLM software. I would not be comfortable with this practice if $\mu$ spanned a wide range.

The R code that I used to produce the figures accompanies the Web version of this article.