Generalized Linear Models Beyond the Exponential Family with Loss Reserve Applications

Gary G. Venter, FCAS, MAAA

Abstract

The formulation of generalized linear models in *Loss Models* by Klugman, Panjer, and Willmot [5] is a bit more general than is often seen, in that the residuals are not restricted to following a member of the exponential family. Some of the distributions this allows have potentially useful applications. The cost is that there is no longer a single form for the likelihood function, so each has to be fit directly. Here the use of loss distributions (frequency, severity, and aggregate) in generalized linear models is addressed, along with a few other possibilities.

Keywords. Loss reserving; regression modeling; generalized linear models.

1 INTRODUCTION

The paradigm of a linear model is multiple regression, where the dependent variables are linear combinations of independent variables plus a residual term, which is from a single mean-zero normal distribution. Generalized linear models, denoted here as GLZ[1], allow nonlinear transforms of the regression mean as well as other forms for the distribution of residuals.

Since many actuarial applications of GLZ are to cross-classified data, such as in a loss development triangle or classification rating plan, a two-dimensional array of independent observations will be assumed, with a typical cell's data denoted as $q_{w,d}$. That is easy to generalize to more dimensions or to a single one.

Klugman, Panjer, and Willmot (2004) [5] provide a fairly general definition of GLZs. To start with, let $\mathbf{z}_{w,d}$ be the row vector of covariate observations for the w, d cell and $\boldsymbol{\beta}$ the column vector of coefficients. Then a GLZ with that distribution models the mean of $q_{w,d}$ as a function η of the linear combination $\mathbf{z}_{w,d}\boldsymbol{\beta}$, where all the other parameters, including $\boldsymbol{\beta}$, are constant across the cells.

It appears that their intention is that η does not take any of the parameters of the

[1] Often GLM is used but with more restrictions on distributional form, typically the exponential family.

distribution as arguments, although this is not explicitly stated. An interesting special case is where η is the identity function, so the mean of $q_{w,d}$ is $\mathbf{z}_{w,d}\boldsymbol{\beta}$. Another key case is where η is exp, so $E[q_w] = \exp(\mathbf{z}_{w,d}\boldsymbol{\beta})$. This is a multiplicative model in which the mean is the product of the exponentiated summands in $\mathbf{z}_{w,d}\boldsymbol{\beta}$.

Standard regression sets the mean $\mathbf{z}_{w,d}\boldsymbol{\beta}$ to the μ of a normal distribution, which has another parameter σ that is constant across the cells. But almost any distribution that has a mean could be reparameterized so that the mean is one of the parameters. This allows virtually any distribution to be used for the residuals. The mean-parameter will be referred to as μ hereafter.

Usual GLM requires the distribution to be from the exponential family. Mildenhall (1999) [7] defines this as a distribution that can be written in the form $f(x;\mu,\phi) = c(x,\phi)/\exp[d(x;\mu)/(2\phi)]$ where $d(x;\mu) = 2w\int_{\mu}^{x} \frac{x-\mu}{V(t)}dt$ for a strictly positive function $V(t)$ and weighting constant w. The tricky part is that μ appears only in the exponent and is constrained in how it combines with ϕ. For any μ, c has to make the integral unity. While quite a few models are possible with this family and various η functions, expanding the universe of distributions leads to other interesting models. Some of the simplicity of exponential models is lost, however.

Standard theory shows the mean of an exponential model is μ and the variance is $\phi V(\mu)/w$. The V function defines the exponential model uniquely. Using w=1 and $V = \mu^j$ with j = 0, 1, 2, 3 gives the normal, Poisson, gamma, and inverse Gaussian distributions, respectively. The ratio of the coefficient of skewness to the coefficient of variation (or CV, which is the standard deviation divided by mean) for these distributions is also 0, 1, 2, 3, respectively. Renshaw (1994) [10] has a formula that implies more generally that skewness/CV is $\mu\partial lnV/\partial\mu$ whenever w=1.

The relationship of variance to mean is one of the issues in selecting a distribution for GLZs. The relationship no longer uniquely defines the distribution, however. For the

normal and *t*-distributions[2] the mean and variance are not related, which could be expressed as the variance being proportional to μ^0. The Poisson has variance proportional to μ^1, and quite a few distributions have variance proportional to μ^2. Other relationships of mean and variance will be discussed below. One advantage of GLZs is that distributions with the same relationship of variance to mean might have different tail properties, including different skewnesses and higher moments, giving more flexibility in fitting models to data.

In linear regression the failure of the observations to match the predictions of constant variance is called heteroscedasticity. Often this occurs because the variance is smaller for smaller observations. In such a case using a distribution with variance proportional to a power of the mean might solve the heteroscedasticity problem. A simple example is the Poisson, where μ is the λ parameter, which then gets set to $\eta(\mathbf{z}_{w,d}\boldsymbol{\beta})$ for each cell and then is the mean and variance of the cell.

Virtually any distribution can be used in a GLZ. Specific examples of frequency, severity, and aggregate loss distributions in GLZs are discussed next, followed by estimation issues and examples from modeling loss development triangles.

2 FREQUENCY DISTRIBUTIONS IN GLZ

For the Poisson in λ, the mean and variance are both $\lambda = \eta(\mathbf{z}_{w,d}\boldsymbol{\beta})$. The negative binomial is more interesting. In the usual parameterization, the variance is a fixed multiple of, but greater than, the mean. Negative binomial distributions are in the $(a,b,0)$ class, which means that for $k>0$, there are values a and b so that probabilities follow the recursive relationship $p_k = (a+b/k)p_{k-1}$. The negative binomial has two positive parameters, r and β, with mean = $r\beta$ and variance = $r\beta(1+\beta)$. Skewness/CV is $1+\beta/(1+\beta)$, which is between 1 and 2. Probabilities start with $p_0 = (1+\beta)^{-r}$ and in the recursion $a = \beta/(1+\beta)$ and $b = (r-1)a$.

There are two simple ways to express the negative binomial mean as a parameter. First, keeping the parameter β, replace r by μ/β, so there are two parameters β and μ and the mean

[2] Having *t*-distributed residuals is one of the many possibilities this formulation of GLZ allows. Also the Laplace, which has exponential tails in both directions from the origin, or the logistic, which is like a heavy-tailed normal, could be used for symmetric residuals.

is μ. The variance $r\beta(1+\beta)$ becomes $\mu(1+\beta)$. In a GLZ the mean is $\mu = \eta(\mathbf{z}_{w,d}\boldsymbol{\beta})$ and the variance is $\eta(\mathbf{z}_{w,d}\boldsymbol{\beta})(1+\beta)$, which is proportional to the mean. On the other hand if you keep r and replace β by μ/r, the parameters are r and μ, and the mean is again μ, but the variance $r\beta(1+\beta)$ is $\mu(1+\mu/r)$, which is quadratic in μ. This form is in the exponential family. Thus depending on how you parameterize the negative binomial, its variance can be either linear or quadratic in the mean.

The parameterization chosen does not make any difference for a single distribution. Suppose for example that X has $r = 3$ and $\beta = 10$ and so mean $\mu=30$ and variance 330. The variance is $\mu(1+\beta)$ in the first formulation and $\mu(1+\mu/r)$ in the second, both of which are 330. A difference comes when modeling other variables while keeping parameters other than μ constant. Suppose Y has mean 100. If β is kept at 10, $\mu(1+\beta) = 1100$, while if r is kept at 3, $\mu(1+\mu/r) = 3433$. The parameterization to choose would be the one that best captures the way the variance grows as the risk size increases. This same idea is applied to severity distributions next.

3 SEVERITY DISTRIBUTIONS IN GLZ

A parameter θ of a distribution of X is a scale parameter if the distribution of a multiple of X is obtained by substituting that multiple of θ into the original distribution. The k^{th} moment of the distribution is then proportional to θ^k. Thus if the mean μ is a scale parameter, the variance is proportional to μ^2.

3.1 Inverse Gaussian

As an example, consider the inverse Gaussian distribution with density

$$ig_1(x;\mu,\alpha) = \sqrt{\frac{\mu}{2\pi\alpha x^3}}e^{\frac{2-x/\mu-\mu/x}{2\alpha}}.$$

Here μ is a scale parameter, with $EX = \mu$ and $VarX = \alpha\mu^2$. However it is more usual to parameterize the inverse Gaussian with $\lambda = \mu/\alpha$, so α is replaced by μ/λ:

$$ig_2(x;\mu,\lambda) = e^{\frac{2-x/\mu-\mu/x}{2\mu/\lambda}}\sqrt{\frac{\lambda}{2\pi x^3}}.$$

Now μ is no longer a scale parameter, even though it is still the mean. The variance is μ^3/λ, and so is proportional to μ^3 instead of μ^2. This is in the exponential family as μ is just in the exponent. Both forms meet the requirements to be GLZs, so either variance assumption can be accommodated. The choice would depend on how the squared deviations from the cell means tend to vary with the means $\eta(\mathbf{z}_{w,d}\boldsymbol{\beta})$. If they seem to grow proportionally to the square of the mean, ig_1 would be indicated, but if they grow with the mean cubed, ig_2 would be preferred.

How the variance relates to the mean is thus not a fundamental feature of the inverse Gaussian, but is a result of how it is parameterized. A characteristic constant of this distribution, not dependent on parameterization, is the ratio of the skewness to the CV. In ig_1, with μ a scale parameter, the third central moment is $3\mu^3\alpha^2$ while it is $3\mu^5/\lambda^2$ in ig_2. Thus in ig_1 the CV is $\alpha^{1/2}$ and the skewness is $3\alpha^{1/2}$, so the ratio is 3. In ig_2 these coefficients are $(\mu/\lambda)^{1/2}$ and $3(\mu/\lambda)^{1/2}$, so the ratio is again 3.

3.2 Gamma

Substituting alternative parameters can be done for other distributions as well. For instance the gamma distribution is usually parameterized $F(x; \theta,\alpha) = \Gamma(x/\theta; \alpha)$ with the incomplete gamma function Γ. This has mean $\alpha\theta$ and variance $\alpha\theta^2$. To get the mean to be a parameter, set $F(x; \mu,\alpha) = \Gamma(x\alpha/\mu; \alpha)$. Then the variance is μ^2/α and μ is still a scale parameter. But other parameterizations are possible. Similarly to the inverse Gaussian, setting $F(x; \mu,\lambda) = \Gamma(x\lambda/\mu^2; \lambda/\mu)$ still gives mean μ but now the variance is μ^3/λ. Other variance functions can be reached by this method. For instance $F(x; \mu,\lambda) = \Gamma[x/(\lambda\mu^p); \mu^{1-p}/\lambda]$ has mean μ and variance $\mu^{1+p}\lambda$. This works for any real p, so the gamma variance can be made to be proportional to any power of the mean, including zero. This will be called the gamma p.

Hewitt (1966) [3] noted that if larger risks were independent sums of small risks, the variance would grow in proportion to the mean. He found in fact that aggregate loss distributions for some insurance risks can be modeled by gamma distributions, and that the gamma variance grows by about $\mu^{1.227}$. This relationship could be modeled by the gamma p with $p = 0.227$.

As with the inverse Gaussian, the ratio of skewness to CV is a characteristic constant of the gamma distribution. With power p, the third central moment is $2\lambda^2\mu^{1+2p}$. This gives skewness of $2\lambda^{0.5}\mu^{0.5p-0.5}$, which is twice the CV, so the ratio is 2 for the gamma regardless of p. Thus an inverse Gaussian is 50% more skewed than the gamma with the same mean and variance.

3.3 Lognormal

The lognormal density can be parameterized as:

$$f(x;\theta,\tau) = \frac{e^{-[\log(x/\theta)]^2/(2\tau)}}{x\sqrt{2\pi\tau}}.$$

Here θ is a scale parameter. The mean is $\theta e^{\tau/2}$ and the variance is $\theta^2 e^{\tau}(e^{\tau}\text{-}1)$. Taking $\alpha = e^{\tau/2}$ and $\mu = \alpha\theta$, the mean and variance are μ and $\mu^2(\alpha^2 - 1)$ and

$$f(x;\mu,\alpha) = \frac{e^{-[\log(\alpha x/\mu)]^2/(4\log\alpha)}}{2x\sqrt{\pi\log\alpha}}.$$

For the lognormal a characteristic constant is the ratio of skewness to CV minus the CV-squared. This is always 3, regardless of parameterization.

The usual parameterization of the lognormal is: $F(x;\mu,\sigma) = N\left(\frac{\ln(x)-\mu}{\sigma}\right)$. This has mean $e^{\mu+\sigma^2/2}$ and variance $e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right)$. Now reparameterize with two parameters m and s:

$$F(x;m,s) = N\left(\frac{\ln\left((x/m)\sqrt{1+s^2/m^2}\right)}{\sqrt{\ln\left(1+s^2/m^2\right)}}\right)$$

It is not hard to see that μ has been replaced by $\ln\left(\frac{m^2}{\sqrt{s^2+m^2}}\right)$ and σ^2 has been replaced by $\ln\left(\frac{s^2+m^2}{m^2}\right)$. Thus e^{μ} is $\frac{m^2}{\sqrt{s^2+m^2}}$ and e^{σ^2} is $\frac{s^2+m^2}{m^2}$. From this it follows that the mean is m and the variance is s^2. This parameterization makes the mean and variance completely unrelated. By the way, skewness is then also a fairly simple function of the

parameters: skewness = $3\frac{s}{m}+\frac{s^3}{m^3}$. As with the gamma, other reparameterizations of the lognormal are possible, and can give any relationship of variance and mean. In fact,

$$F(x; m, s, p) = N\left(\frac{\ln\left((x/m)\sqrt{1+s^2m^{p-2}}\right)}{\sqrt{\ln\left(1+s^2m^{p-2}\right)}}\right)$$

has mean m, variance s^2m^p, and skewness $3t+t^3$, where $t = sm^{p/2-1}$. Here μ has been replaced by $\ln\left(\frac{m}{\sqrt{1+s^2m^{p-2}}}\right)$ and σ^2 by $\ln(1+s^2m^{p-2})$.

3.4 Pareto

The Pareto is another interesting case. Consider $F(x; \theta,\alpha) = 1 - (1+x/\theta)^{-\alpha}$. This has mean $\theta/(\alpha-1)$. Taking $\mu = (\alpha-1)\theta$ gives $F(x; \mu,\alpha) = 1 - (1+x/(\mu\ \alpha-\mu))^{-\alpha}$. This has mean μ and variance $\mu^2/(\alpha-2)$ if $\alpha > 2$. But if $\alpha \le 1$ this does not work, as the mean does not exist. There does not seem to be any reason not to extend the GLZs to this case. Perhaps the easiest way to do this is to model $\theta_{w,d}$ as $\eta(\mathbf{z}_{w,d}\boldsymbol{\beta})$ for each cell. Or the median $m = \theta(2^{1/\alpha} - 1)$ could be the parameter modeled, by setting $F(x; m,\alpha) = 1-(1+x(2^{1/\alpha}-1)/m)^{-\alpha}$, with $m = \eta(\mathbf{z}_{w,d}\boldsymbol{\beta})$. This is median regression in the GLZ framework.

The skewness for the gamma, inverse Gaussian and lognormal distributions can be expressed as 2CV, 3CV, and $3CV+CV^3$, respectively. For the Pareto, if the skewness exists, CV^2 is in the range (1,3). Then the skewness is $\frac{2}{CV}\frac{\alpha+1}{\alpha-3} = 2CV\frac{3-CV^{-2}}{3-CV^2}$. This is less than the lognormal skewness when $CV^2 < 2$ and less than the inverse Gaussian skewness when $CV^2 < 0.5+\sqrt{11/12} \approx 1.4574$. This illustrates the different tail possibilities for GLZs with the same means and variances.

3.5 Origin Shifting

Severity distributions have their support on the positive reals, so all fitted values have to be positive. Frequency and aggregate distributions extend the support to include zero, but not negative values. However, any of the positive distributions can be location shifted to allow the possibility of negative values or even negative means. For instance, the shifted

gamma has $F(x) = \Gamma[(x-b)/\theta, \alpha]$, with mean $b+\alpha\theta$ and variance $\alpha\theta^2$. Making the mean a parameter gives the distribution $F(x) = \Gamma[(\alpha(x-b)/(\mu-b), \alpha]$. The variance is then $(\mu-b)^2/\alpha$, which is still quadratic in μ.

4 AGGREGATE DISTRIBUTIONS IN GLZ

Aggregate distributions can be especially useful for residuals that are continuous on the positive reals but also could take a positive probability at zero. This is often seen out in late lags of a development triangle, for example.

4.1 Poisson-Gamma Aggregates

An example of an aggregate loss model in the exponential family is the Tweedie distribution. This starts by combining a gamma severity in α and θ that has mean $\alpha\theta$ and variance $\alpha\theta^2$ with a Poisson frequency in λ. Then the aggregate distribution has mean $\mu = \lambda\alpha\theta$ and variance $= \lambda\alpha\theta^2(\alpha+1) = \mu\theta(\alpha+1)$. Since this can also be written as $\lambda(\alpha\theta)^2(1/\alpha+1)$, it is clear that the variance is linear in the frequency mean and quadratic in the severity mean.

If the restriction $\lambda = k(\alpha\theta)^\alpha$ is imposed, then $\mu = k(\alpha\theta)^{\alpha+1}$, and the variance is $k\alpha^{\alpha+1}\theta^{\alpha+2}(1+\alpha)$, or $\mu^{1+1/(\alpha+1)}(1+1/\alpha)k^{-1/(\alpha+1)}$. This is the Tweedie distribution. The variance is proportional to a power of the mean between 1 and 2, which is often realistic for sums of claims. The link between frequency and severity is problematic, however. It would seem unusual for the observations with the smaller number of claims to also have the smaller claim sizes.

Kaas (2005) [4] expresses the Tweedie by replacing the three parameters λ, α, θ of the Poisson-Gamma with three others μ, ψ, and p by the formulas:

$$\lambda = \mu^{2-p}/[\psi(2-p)] \qquad \alpha = (2-p)/(p-1) \qquad \theta = \psi(p-1)\mu^{p-1}$$

This looks like a 3 for 3 swap of parameters, so it is not clear that a relationship between the frequency and severity means has been imposed. But $(\alpha\theta)^\alpha$ in this notation is:

$$(\alpha\theta)^\alpha = \lambda[\psi(2-p)]^{1/(p-1)}.$$

Thus taking $k = [\psi(2-p)]^{1/(1-p)}$ gives $\lambda = k(\alpha\theta)^{\alpha}$, which is the restriction originally imposed above. This k is not a function of μ and can also replace ψ by $\psi = k^{1-p}/(2-p)$. This gives a parameterization of the Tweedie in terms of k, p, and μ:

$$\lambda = \mu(\mu/k)^{1-p} \qquad \alpha = (2-p)/(p-1) \qquad \theta = (\mu/k)^{p-1}/\alpha$$

The mean is still μ, the frequency mean is k times the severity mean raised to the power $(2-p)/(p-1)$, and the aggregate variance is now $\mu^p k^{1-p}/(2-p)$. Since p is $(\alpha+2)/(\alpha+1)$, it is between 1 and 2. The parameters are a bit simpler than Kaas' but the variance is more complicated than his $\psi\mu^p$. In any case skewness/CV is p, consistent with Renshaw's formula.

Not requiring the exponential family form gives other possibilities. Without imposing any relationship between frequency and severity, as noted above, the Poisson-gamma can be parameterized with mean μ and variance $\mu\theta(\alpha+1)$. This has replaced λ with $\mu/(\alpha\theta)$. A somewhat different relationship between frequency and severity can be established by setting $\lambda = (\alpha\theta)^k$. This gives mean $\mu = (\alpha\theta)^{k+1}$ and variance $(\alpha\theta)^{k+2}(1+1/\alpha) = \mu^{(k+2)/(k+1)}(1+1/\alpha)$, which is again proportional to a power of the mean between 1 and 2.

4.2 Poisson-Normal

A limiting case is the Poisson-normal. This has a point mass at zero but could have some negative observations. For the normal in m and s^2 it has mean $\mu = \lambda m$, variance $\lambda(m^2+s^2) = \mu m[1+(s/m)^2]$ and skewness $(1+3CV^2)\lambda^{-1/2}(1+CV^2)^{-1.5}$. Fixing m and s and setting $\lambda_{w,d}$ to $\mu_{w,d}/m$ makes the variance proportional to the mean. Another possibility is to make λ and s constant and set $m_{w,d}$ to $\mu_{w,d}/\lambda$. Then the variance of each cell is $\mu_{w,d}^2/\lambda + \lambda s^2$. This is quadratic in $\mu_{w,d}$ and any $\mu_{w,d}$ can be negative. This is possible for the normal regression as well, but for the Poisson-normal, homoscedasticity is not required (or possible).

4.3 Poisson-Constant Severity Aggregates

The simplest aggregate loss distribution is probably Poisson frequency with a constant severity, called the PCS distribution. If θ is the severity, a cell with frequency λ has mean $\theta\lambda = \eta(\mathbf{z}_{w,d}\boldsymbol{\beta})$ and variance $\theta^2\lambda = \theta\eta(\mathbf{z}_{w,d}\boldsymbol{\beta})$. This is sometimes called the over-dispersed Poisson distribution, but PCS may be more descriptive, especially if $\theta < 1$. Some authors define the over-dispersed Poisson more broadly as any distribution in the exponential family

for which the variance is proportional to the mean. But by uniqueness properties of the exponential family the PCS is the only such distribution, and so is the unique over-dispersed Poisson.

If X is the total loss random variable, X/θ is Poisson in $\lambda = EX/\theta = \mu/\theta$. Thus $\Pr(X/\theta = n) = e^{-\mu/\theta}(\mu/\theta)^n/n!$. For $x = \theta n$, $\Pr(X=x) = e^{-\mu/\theta}(\mu/\theta)^{x/\theta}/(x/\theta)!$. If x is not an integer multiple of θ, $\Pr(X=x) = 0$. If μ is modeled by covariates and parameters, say $\mu_{w,d} = U_w g_d$, with θ fixed, then an observation of $X_{w,d}$, say $q_{w,d}$, with $q_{w,d}/\theta$ a non-negative integer, has $\Pr(X_{w,d} = q_{w,d}) = p(q_{w,d}) = e^{-\mu_{w,d}/\theta}(\mu_{w,d}/\theta)^{q_{w,d}/\theta}/(q_{w,d}/\theta)!$, and $p(q_{w,d})$ is zero otherwise. The PCS is a discrete distribution with positive probability only at integer multiples of θ. By its uniqueness, there is no continuous over-dispersed Poisson distribution in the exponential family. Thus over-dispersed Poisson probabilities are always zero except at integer multiples of θ.

A continuous analogue of the PCS is discussed in Mack (2002)[3] [6]. This can be described as a zero-modified continuous scaled Poisson, or ZMCSP. To specify it, start by using $p(x)/\theta$ as a density on the positive reals, extending the factorial by the gamma function, i.e., defining $a! \equiv \Gamma(1+a)$. But this density gives total probability above unity. Mack's solution is to reduce the probability mass at zero.

The ZMCSP is defined by the density $f(x;\mu,\theta) = e^{-\mu/\theta}(\mu/\theta)^{x/\theta}/[\theta(x/\theta)!]$ for $x > 0$ and by setting the point mass at $x = 0$ enough to make the total probability 1. To see how much probability is needed at 0, define the function $\text{pois}(x,\lambda) = \lambda^x e^{-\lambda}/x!$ and the function $\text{zm}(\lambda) = 1 - \int_{0+}^{\infty} \text{pois}(x,\lambda)dx$. Then with a change of variable in $f(x)$ to $y = x/\theta$ and defining $\lambda = \mu/\theta$, it is easy to see that $\int_0^{\infty} f(x;\mu,\theta)dx$ is $1 - \text{zm}(\lambda)$. Thus the point mass needed at zero is $\text{zm}(\mu/\theta)$. The function $\text{zm}(\lambda)$ is less than the Poisson's point mass of $e^{-\lambda}$ but is strictly positive.

There is an extra θ in the denominator of f that is not in p, but that will not affect the MLE of μ or the components of μ if μ is a function of covariates. This is interesting because setting $\mu_{w,d} = U_w g_d$ in the PCS and estimating by MLE is known to give the chain-ladder

[3] Chapter 1.3.7 [6].

reserve estimates. Since the estimates of U_w and g_d for Mack's ZMCSP will be the same as for the PCS (as long as there are not any zero observations), this looks like it extends the match of the chain ladder to the continuous case - no longer requiring that all cells in the triangle are integer multiples of θ. It turns out however that this is approximately but not exactly so.

The divergence arises from the fact that the ZMCSP expected value is not exactly μ. Integrating $xf(x)$ shows that the mean is actually:

$$EX = \mu[1 - \text{zm}(\mu/\theta) + \int_{-1}^{0} \text{pois}(x,\mu/\theta)dx].$$

This is greater than μ, but not by much, unless λ is small, as Table 1 shows. Since the function of μ needed to produce the mean depends on the parameters of the distribution, the ZMCSP is probably not a GLZ. As with the Pareto with infinite mean, extending the definition of GLZ a bit to include linear modeling of a parameter that is not the mean may make sense. Whether or not this is considered a GLZ, it is still a useful model.

The variance is a bit less than θμ for small values of λ. Integrating $x^2f(x)$ shows that $EX^2 = \theta^2\lambda \int_{0+}^{\infty} \text{pois}(x-1,\lambda)xdx$. For large values of λ the integral is λ+1, but it is different for smaller λ.

Table 1: Point mass and moment adjustment by λ

λ = μ/θ	zm(μ/θ)	$EX/\mu - 1$	$EX^2/[\theta^2\lambda(\lambda+1)] - 1$	$\text{Var}/\theta^2\lambda - 1$
0.2	.48628	.33861	.03976	-0.11066
1	.16619	.03291	–8.73e-04	-0.06865
5	.00216	9.43e-05	-3.75e-06	-0.00097
25	3.19e-12	1.96e-14	-7.00e-13	-1.9E-11

In a recent study of a fairly noisy runoff triangle, μ/θ was less than two for just one observation and less than five for five observations, out of 55. Thus, a few small observations would have fitted means a percent or two different from the chain ladder's. While the noted match of the PCS and chain-ladder reserve estimates holds exactly only when all probability is concentrated on integer multiples of θ, the ZMCSP comes close to having this relationship in the continuous case.

4.4 Geometric – Exponential

The geometric frequency distribution can be described with a parameter α by $p_k = \alpha(1 - \alpha)^k$ for $k \geq 0$. This has mean $(1 - \alpha)/\alpha$ and variance $(1 - \alpha)/\alpha^2$, which is higher than the mean. With an exponential severity in mean θ, the aggregate distribution has mean $\theta(1 - \alpha)/\alpha$ and variance $\theta^2(1 - \alpha^2)/\alpha^2$. The aggregate survival function is known[4] to be $S(x) = (1 - \alpha)e^{-x\alpha/\theta}$. Both the frequency and aggregate distributions have a point mass of α at 0.

Either α or θ can be replaced by the mean μ, but probably keeping a constant θ would be useful more often. This replaces α by $\theta/(\mu+\theta)$. Thus when μ is higher, the probability α of an observation of zero is lower, which would make sense in many cases. The aggregate mean and variance become μ and $\mu(\mu+2\theta)$ with survival function $S(x) = \mu/(\mu+\theta)e^{-x/[\mu+\theta]}$. The variance is quadratic in the mean but with the linear term it increases more slowly than μ^2. For MLE the aggregate density is $f(x) = \mu/(\mu+\theta)^2 e^{-x/[\mu+\theta]}$ for $x > 0$ and $p_0 = \theta/(\mu+\theta)$.

5 ESTIMATION ISSUES

Key to estimation is having an efficient optimizer to estimate the likelihood function including the covariates. Advances in computing power and the availability of optimization algorithms, even as spreadsheet add-ins, is what makes it possible to go beyond the exponential family and to use full MLE estimation.

The modified distributions like gamma *p* and lognormal basically substitute formulas for the usual parameters. For example in the gamma *p*, $F(x; \mu,\lambda) = \Gamma[x/(\lambda\mu^p); \mu^{1-p}/\lambda]$ can be written as $F(x) = \Gamma(x/\theta; \alpha)$ with $\theta = \lambda\mu^p$ and $\alpha = \mu^{1-p}/\lambda$. Thus a routine that searches for optimal gamma parameters can be used to estimate the gamma *p* by first expressing the gamma parameters in terms of λ, μ, and *p* and then searching for the best values for these three parameters. Since μ will be a function of covariates involving several parameters, this is the part where efficient algorithms comes in.

As long as there are no zero observations the ZMCSP loglikelihood function is

[4] See *Loss Models* [5], page 154.

$l = \sum\left(\frac{q_{w,d}}{\theta}\ln\frac{\mu_{w,d}}{\theta} - \frac{\mu_{w,d}}{\theta} - \ln\left(\frac{q_{w,d}}{\theta}!\right) - \ln(\theta)\right)$. The last two terms in the sum can be omitted when maximizing for μ. In fact the function to maximize can be reduced to $l^* = \sum\left(q_{w,d}\ln\mu_{w,d} - \mu_{w,d}\right)$. Taking the derivative shows that this is maximized when $0 = \sum\left(\frac{q_{w,d}}{\mu_{w,d}} - 1\right)$. Thus the average relative error should be zero. If μ is a function of covariates and the vector β is being estimated, the derivative of l^* wrt the j^{th} element of β, β_j, gives the n equations $0 = \sum\frac{\partial\mu_{w,d}}{\partial\beta_j}\left(\frac{q_{w,d}}{\mu_{w,d}} - 1\right)$. This could be considered a series of weighted average relative errors, all of which should be 0. After finding the estimates of the β_j, the likelihood can be maximized for θ. The Poisson is analogous to the normal distribution case where the loglikelihood reduces to minimizing $\sum\left(q_{w,d} - \mu_{w,d}\right)^2$. This gives the n equations $0 = \sum\frac{\partial\mu_{w,d}}{\partial\beta_j}\left(q_{w,d} - \mu_{w,d}\right)$. Here the weighted average errors should be 0.

In non-parametric estimation, it is common to adopt the criterion of minimizing the sum of the squared errors, regardless of distribution. This treats a fixed squared error in any observation as equally bad – basically incorporating a constant variance assumption. This reduces to the normal distribution when in the exponential family, so minimizing squared error is a normal non-parametric approach. It sets the sum of weighted errors to 0. This is called unbiased, which sounds like something you want to be, but is not always that important.

If the same weighted relative error is equally bad across observations this is more of a Poisson assumption. This could also be used in a non-parametric context, where the weighted sums of relative errors are set to 0. This could be done without assuming the form of the distribution, so could be a Poisson non-parametric approach. The reasoning above shows that this results from finding the parameters that minimize $\sum\left(fitted - actual\ln fitted\right)$. This forces the actual/fitted toward 1.

For the Poisson-gamma aggregate and its special cases (Tweedie, etc.) the density for the

likelihood function can be calculated by inverting the characteristic function $\varphi(t) = \exp[-1+\lambda(1 - it\theta)^{-\alpha}]$. Mong (1980) [9] worked out a purely real integral for this in terms of λ, α, θ and the aggregate standard deviation σ:

$$f(x) = \frac{1}{\sigma\pi}\int_0^\infty e^{\lambda j(t)} \cos\left[\frac{xt}{\sigma} - \lambda k(t)\right]dt \text{, where } j(t) = \delta(t)\cos[\rho(t)]-1 \text{ ; } k(t) = \delta(t)\sin[\rho(t)]$$

$\delta(t) = [1+(t\theta/\sigma)^2]^{-\alpha/2}$; $\rho(t) = \alpha\tan^{-1}(t\theta/\sigma)$. The scaling by σ is probably done for numerical efficiency. With covariates, $\mu/\theta\alpha$ could replace λ in the characteristic function and its inversion. For the Tweedie it is also possible to express the density using an infinite sum, as in Clark and Thayer (2004) [1].

The gamma characteristic function is $\varphi_\Gamma(t) = (1 - it\theta)^{-\alpha}$, and $\varphi_\Gamma(t/\sigma) - 1 = j(t) + ik(t)$. For the normal distribution in m and s^2 the characteristic function is $\varphi_N(t) = \exp(itm - 0.5(st)^2)$. Scaling by s instead of σ gives $\varphi_N(t/s) - 1 = j(t) + ik(t)$ where $j(t) = \exp(-0.5t^2)\cos(tm/s) - 1$ and $k(t) = \exp(-0.5t^2)\sin(tm/s)$. These can be used in the integral above to give the Poisson-normal density if σ is replaced by s.

Mong's comments are: "(The) formula and its consequent computations may seem complex in the form shown above. However, the implementation is quite simple. Any standard numerical integration technique would handle the computation effectively; for example, the extended Simpson's rule is adequate to calculate the integration and is easy to code in any scientific programming language."

The extended Simpson's rule breaks down a finite range of integration into $2n$ intervals of length h, with $2n+1$ endpoints x_0, …, x_{2n}. The function to be integrated is evaluated at each of the $2n+1$ points and multiplied by h. Then these are weighted by the following factors and summed: x_0 and x_{2n} get weight 1/3; odd points x_1, x_3, …, x_{2n-1} get weight 4/3; even points x_2, …, x_{2n-2} get weight 2/3.

Figure 1: Integrand for Poisson-gamma density

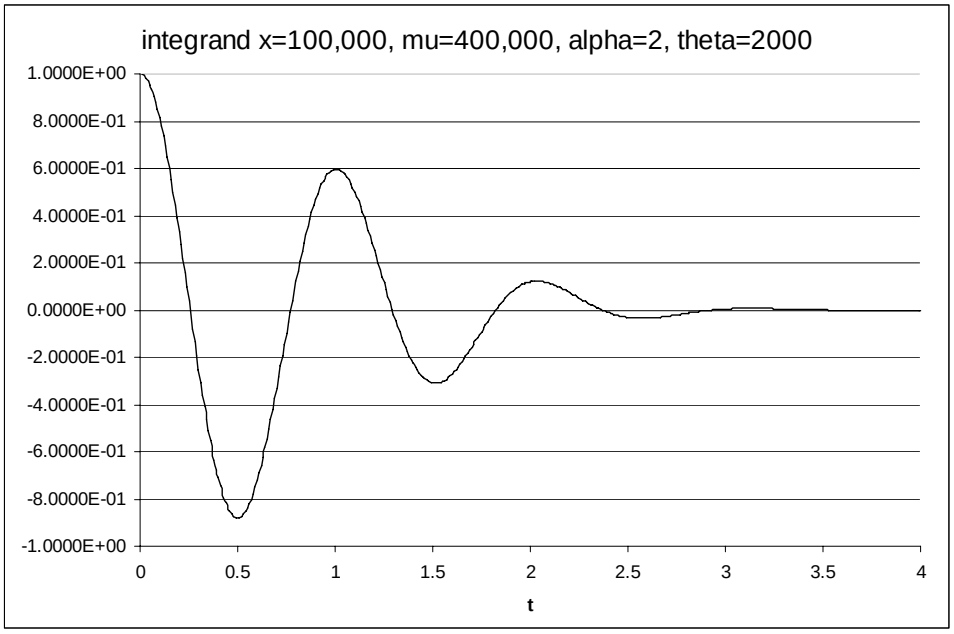


Figure 1 shows an example of the integrand for the Poisson-gamma density. This is for an x that is more than six standard deviations below the mean for a positively skewed distribution, so the integrated probability is low (7.5e-19). This makes the integration a bit more difficult as the dampening cycles have to get quite small before it stabilizes. However this occurred by about $t = 10$. Less remote probabilities have cycles that damp out more quickly.

There is a problem with this integral, however. The integration for $f(x)$ does not converge[5]! For both the gamma and normal severities, as t gets large $j(t) \to -1$ and $k(t) \to 0$. Thus the integrand becomes $e^{-\lambda}\cos(xt/\sigma)/(x\sigma)$, which fluctuates and does not go to 0. If λ is sufficiently large, this fluctuation is well beyond any reasonable degree of accuracy, and so is not a problem. Otherwise an alternative is to use the inversion formula for the distribution function to calculate $[F(x+\varepsilon) - F(x-\varepsilon)]/2\varepsilon$ for some appropriate ε, perhaps ½. According to Mong that inversion is: $F(x) = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \frac{e^{\lambda j(t)}}{t} \sin\left[\frac{xt}{\sigma} - \lambda k(t)\right] dt$, which does converge.

[5] My colleague John Major pointed this out.

6 DEVELOPMENT FACTOR EXAMPLE

Venter (2007) [12] fit the development triangle in Table 2 by a regression model for the incremental losses at lags 1 and above. The independent variables were the cumulative losses at lags 0 through 4, a dummy variable equal to 1 for the 3rd diagonal and 0 elsewhere, a dummy variable equal to 1 on the 4th, 7th, and 9th diagonals, -1 on the 10th diagonal, and 0 elsewhere, and a constant term. The diagonals are numbered starting at 0, so the 3rd is the one beginning with 7,888 and the 10th starts with 19,373. The cumulative loss independent variables are set to 0 for incremental losses that are not in the immediately following column.

Table 2: Cumulative Loss Development Triangle

Lag0	Lag1	Lag2	Lag3	Lag4	Lag5	Lag6	Lag7	Lag8	Lag9	Lag10	Lag11
11,305	30,210	47,683	57,904	61,235	63,907	64,599	65,744	66,488	66,599	66,640	66,652
8,828	22,781	34,286	41,954	44,897	45,981	46,670	46,849	47,864	48,090	48,105	48,721
8,271	23,595	32,968	44,684	50,318	52,940	53,791	54,172	54,188	54,216	54,775	
7,888	19,830	31,629	38,444	43,287	46,032	47,411	47,677	48,486	48,498		
8,529	23,835	35,778	45,238	51,336	53,574	54,067	54,203	54,214			
10,459	27,331	39,999	49,198	52,723	53,750	54,674	55,864				
8,178	20,205	32,354	38,592	43,223	44,142	44,577					
10,364	27,878	40,943	53,394	59,559	60,940						
11,855	32,505	55,758	64,933	75,244							
17,133	45,893	66,077	78,951								
19,373	50,464	75,584									
18,433	47,564										
20,640											

This is a loss development model with a constant term and calendar-year adjustments up through lag 5, but for lags 6 and beyond the constant term and the calendar-year adjustments operate but there are no development factors. The late development appears to be random in time and not dependent on the level of the accident year. There are heteroscedasticity issues, however. The smaller incremental losses at the end tend to have lower residuals – which actually seems desirable. Also the 0 to 1 development factor fits unreasonably well, so the residuals are also lower for the large increments at lag 1.

To address these issues, the same model was fit using Mack's ZMCSP distribution and the gamma p, where p was -0.29. The other parameters, negative loglikelihood, and AICc/2 are shown in Table 3.

Table 3: Parameters and fit statistics

	lag0	**lag1**	**lag2**	**lag3**	**lag4**	**diag3**	**4+-10**	**const**	**θ,λ,σ**	**-lnL**	**AICc/2**
ZMCPS	1.618	0.508	0.223	0.103	0.026	-2072	107.1	487.9	306.1	637.8	646.9
Gamma *p*	1.624	0.504	0.217	0.102	0.027	-1922	132.0	499.8	3,969.0	630.3	642.0
Normal	1.601	0.499	0.211	0.102	0.021	-1832	801.6	527.8	1,387.7	662.2	671.2

For N observations and p parameters, taking half of the small sample AIC, denoted AICc, penalizes the negative loglikelihood by $Np/(N - p - 1)$. For small samples ($N < 40p$) this is growing in popularity as the best way to penalize for extra parameters. Usually all parameters are penalized but for comparing fits maybe parameters that do not affect the fit can be ignored. Here for the normal and ZMCSP, p was set to 8, as θ and σ do not affect the fit. However for gamma p it was set to 10, as λ and p do. Still it gave the best AICc. N is 77 for this data.

The fit is clearly worse for the normal regression, reflecting the heteroscedasticity issue. The variance for the gamma p is $\mu^{0.71}$. Usually a power less than 1 is not anticipated, thinking of losses coming from a compound frequency-severity distribution. The abnormally good fit for the 0 to 1 factor, which has the largest observations, may be pushing the power down. The regression coefficients are quite similar for all the distributions, reflecting the common wisdom that heteroscedasticity does not greatly distort regression estimates. The distribution of possible results will vary among the distributions, however.

Figures 2 and 3 compare PP Plots for the normal and gamma p fits. The gamma p looks more reasonable. Figures 4 and 5 look at standardized residuals vs. fitted for the two distributions. They both look positively skewed, which they should be for gamma p but not for normal. Also the normal extremes are more extreme. The small fitted values have standardized residuals more like the other values for the gamma p, but not for the normal. Overall the gamma p seems to fit better.

Figure 2

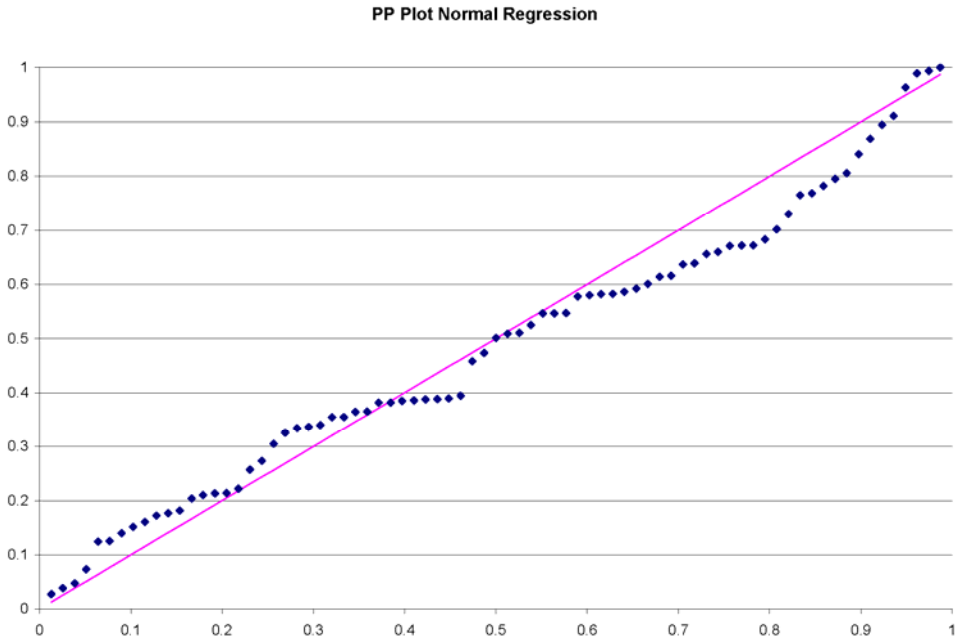


Figure 3

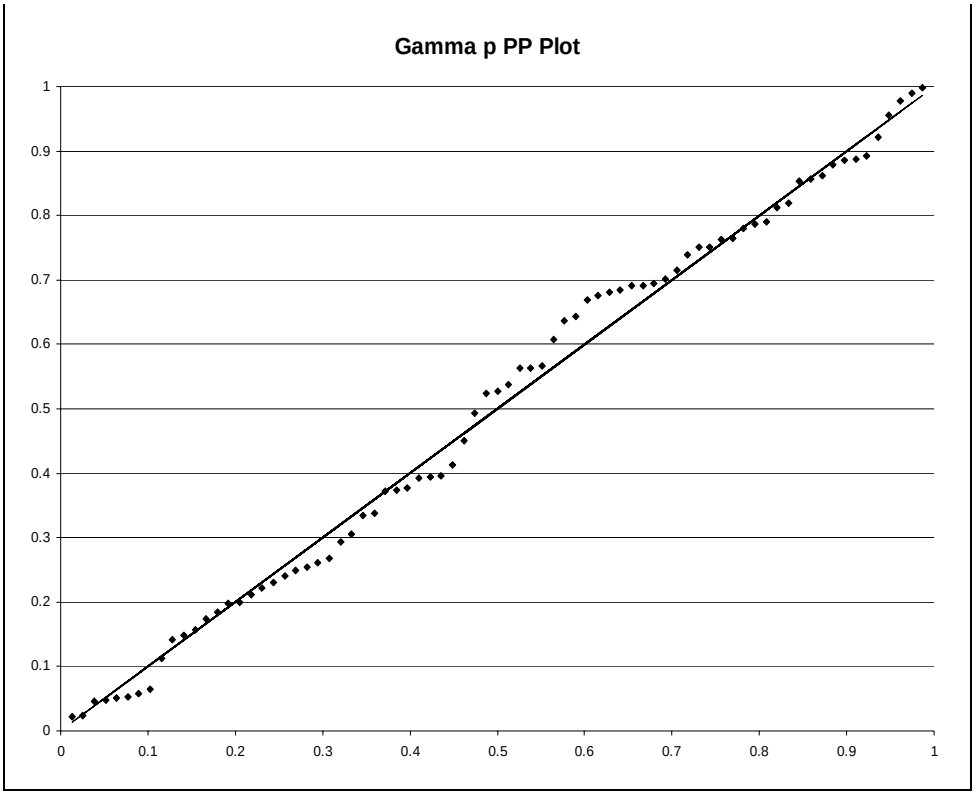


Figure 4

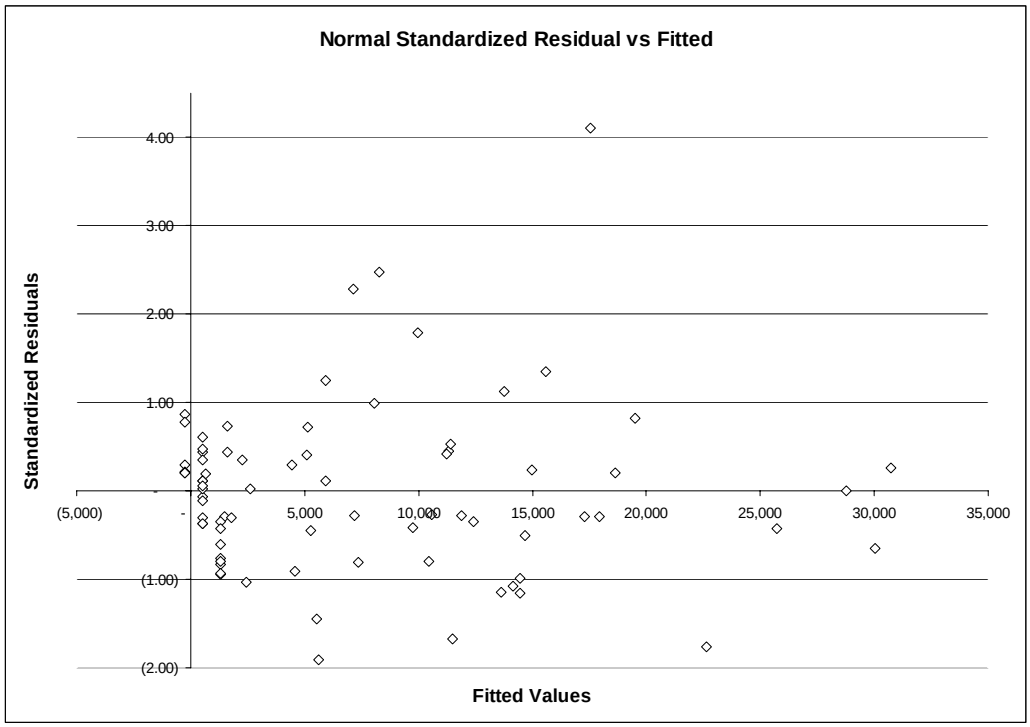
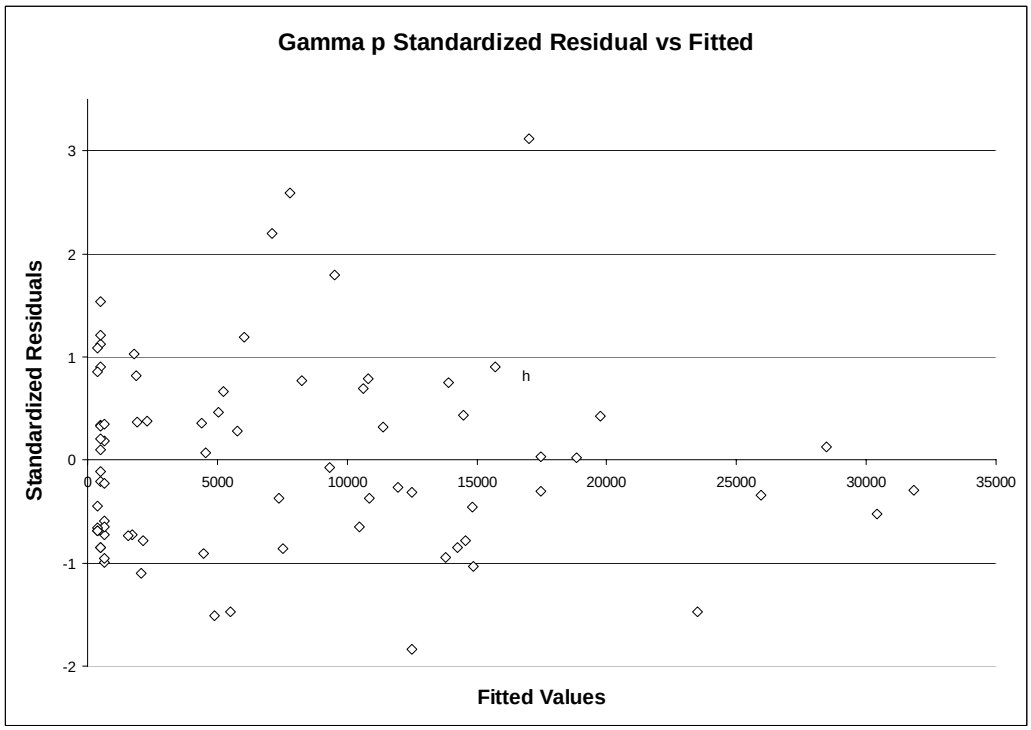


Figure 5



7 MULTIPLICATIVE MODEL ISSUES

Multiplicative fixed-effects models can be treated in the GLZ framework. Take the case where $\mu_{w,d} = Eq_{w,d} = U_w g_d h_{w+d}$. The covariates are 0, 1 dummies picking out which factors apply to each cell, and the vector of coefficients $\boldsymbol{\beta}$ is the log of all the accident year factors U_w followed by the log of all the delay factors g_d followed by the log of all the calendar year factors h_{w+d} in the model. Let $\mathbf{z_{w,d}}$ be the vector that has zero in all positions except for ones for the positions of the w^{th} row, d^{th} column and $w+d^{\text{th}}$ diagonal. Then $\eta(\mathbf{z_{w,d}}\boldsymbol{\beta}) = \exp(\mathbf{z_{w,d}}\boldsymbol{\beta})$ is $E\mu_{w,d}$. This can be used in any of the distributions discussed above. However the factors all have to be positive to take the logs, even though some observations can be negative with the right choice of distribution around the mean. However, if negative means are needed for some columns, $\mu_{w,d} = Eq_{w,d} = U_w g_d h_{w+d}$ with some negative g's can be used directly as the mean of any of the distributions discussed. This could be fit by MLE, but it would not really be considered a linear model any more, unless $\boldsymbol{\beta}$ is allowed to have complex coefficients that become negative reals when exponentiated. The line between GLZ and truly non-linear models is thus a bit imprecise, but the labeling is not really very important anyway.

Fu and Wu (2005) [2] provide an iterative scheme, using constants labeled here as r and s, that can in some cases help in the estimation of multiplicative models. The Fu-Wu iteration for the row-column model can be expressed as[6]:

$$g_d = \left\{ \sum_{w=0}^{n-d} U_w^{r-s} q_{w,d}^{s} \Big/ \sum_{w=0}^{n-d} U_w^{r} \right\}^{1/s} \text{ and } U_w = \left\{ \sum_{d=0}^{n-w} g_d^{r-s} q_{w,d}^{s} \Big/ \sum_{d=0}^{n-w} g_d^{r} \right\}^{1/s}.$$

The idea is to start with reasonable guesses for the U's and then alternatively apply the two formulas to get new g's and U's until they converge. Often this iteration gives the MLE for some model. For instance, taking $r = 2$ and $s = 1$ gives the normal regression. The case $r=s=1$ gives the estimate where $q_{w,d}$ is Poisson in $U_w g_d$. Both of these cases work fine if some column of q's tends to be negative and so its mean g is as well. Mildenhall (2005) [8] shows that there is a model for each r and s for which this iteration gives a reasonable estimate. The cases s=1, r = -1, 0, 1, and 2 are the inverse Gaussian, gamma, Poisson, and normal

[6] They also include weighting factors that here are set to unity.

distributions, respectively, and the estimates are MLE for the β's if the other parameters are known or do not affect the estimates of the β's.

With arbitrary *s* the power transforms of these distributions are realized. Taking $r=0$ gives the transformed gamma or inverse transformed gamma, depending on the sign of s, and so a wide range of distribution shapes. If $1 < r < 2$ and $s = 1$, the Tweedie with $p = r$ is produced. If p and ψ are known, the iteration gives the MLE for the β's. This could be done within an optimization routine that is looking for the MLE values for p and ψ, so would only require a routine that works for two variables.

For the multiplicative models with diagonal factors $E[q_{w,d}] = U_w g_d h_{w+d}$, the Fu-Wu iterative estimates become:

$$g_d = \left[\sum_{w=0}^{n-d} (U_w h_{w+d})^{r-s} q_{w,d}^s \Big/ \sum_{w=0}^{n-d} (U_w h_{w+d})^r \right]^{1/s},$$

$$U_w = \left[\sum_{d=0}^{n-w} (g_d h_{w+d})^{r-s} q_{w,d}^s \Big/ \sum_{d=0}^{n-w} (g_d h_{w+d})^r \right]^{1/s}, \text{ and}$$

$$h_j = \left[\sum_{w+d=j} (U_w g_d)^{r-s} q_{w,d}^s \Big/ \sum_{w+d=j} (U_w g_d)^r \right]^{1/s}.$$

8 MULTIPLICATIVE MODEL EXAMPLE

Table 4 is a development triangle from Taylor-Ashe (1983) [11]. Venter (2007) [12] fit a form of the PCS multiplicative effects model to this data. Each cell $\mu_{w,d}$ was set to the product of row, column, and diagonal effects, but some parameters are used more than once. Accident year 0, a low year, gets its own parameter U_0. Accident year 7 also gets its own parameter U_7 as it is high. All the other years get the same parameter U_a, except year 6 which is a transition and gets the average of U_a and U_7. Thus, there are three accident-year parameters.

The years are divided into high and low payment years with parameters g_a and g_b for fraction of total loss paid in the year. Delay 0 is a low year as payments start slowly. Delays 1, 2, and 3 are the higher payment lags and all get g_b. Delays 5, 6, 7, and 8 are again low getting

g_a. Delay 4 is a transition and gets the average of g_a and g_b. Finally delay 9 gets the rest, i.e., $1 - 5.5g_a - 3.5g_b$. Thus there are only two delay parameters. Three of the diagonals were modeled as high or low, getting factors $1+c$ or $1-c$. The 7th diagonal is low and the 4th and 6th are high. Thus, only one diagonal parameter c is used. The diagonals are numbered from 0, so the 7th starts with 359,480.

Table 4: Incremental triangle Taylor-Ashe (1983) [11]

Lag 0	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9
357,848	766,940	610,542	482,940	527,326	574,398	146,342	139,950	227,229	67,948
352,118	884,021	933,894	1,183,289	445,745	320,996	527,804	266,172	425,046	
290,507	1,001,799	926,219	1,016,654	750,816	146,923	495,992	280,405		
310,608	1,108,250	776,189	1,562,400	272,482	352,053	206,286			
443,160	693,190	991,983	769,488	504,851	470,639				
396,132	937,085	847,498	805,037	705,960					
440,832	847,631	1,131,398	1,063,269						
359,480	1,061,648	1,443,370							
376,686	986,608								
344,014									

Fitting the PCS is done by maximizing $l^* = \sum \left(q_{w,d} \ln \mu_{w,d} - \mu_{w,d} \right)$, where $\mu_{w,d} = U_w g_d h_{w+d}$. This pretends that every observation $q_{w,d}$ is a multiple of θ, as in fact the PCS probability is zero otherwise. This is the same function to be maximized for fitting the ZMCSP, which does not require observations to be multiples of θ. Thus, the row, column, and diagonal parameters are the same for both models. The difference is that θ is fit by an ad hoc method for the PCS and by MLE for ZMCSP. The likelihood function is $l = \sum \left(\frac{q_{w,d}}{\theta} \ln \frac{\mu_{w,d}}{\theta} - \frac{\mu_{w,d}}{\theta} - \ln\left(\frac{q_{w,d}}{\theta}! \right) - \ln(\theta) \right)$, and now θ is the only parameter needed to maximize over. The MLE estimate of θ is 30,892. Estimating it by a moments method $\hat{\theta} = \frac{1}{N-p} \sum_{w,d} \frac{\left(q_{w,d} - U_w g_d \right)^2}{U_w g_d}$ gives 37,184.

Just changing θ makes a difference in the estimated runoff distribution and parameter errors. The estimated parameters and their PCS standard errors from the information matrix with the moment and MLE θ's are in Table 5. The runoff variance separated into process

and parameter is in Table 6.

Table 5: Parameter se's with two estimates of θ

Parameter	U_0	U_7	U_a	g_a	g_b	c
Estimate	3,810,000	7,113,775	5,151,180	0.067875	0.173958	0.198533
se 37,184	372,849	698,091	220,508	0.003431	0.005641	0.056896
se 30,892	339,846	636,298	200,989	0.003127	0.005142	0.051860

Table 6: Runoff Variance with two estimates of θ

Model	**Moment 37,184**	**MLE 30,892**
Parameter Variance	1,103,569,529,544	916,846,252,340
Process Variance	718,924,545,072	597,282,959,722
Total Variance	1,822,494,074,616	1,514,129,212,061
Parameter Std Dev	1,050,509	957,521
Process Std Dev	847,894	772,841
Standard Deviation	1,349,998	1,230,500

So far this is all from keeping the PCS framework and replacing the estimated θ from the moment method by that from MLE from ZMCSP. The ability to estimate θ by MLE is actually the main difference between the two distributional assumptions. In this case the MLE θ is quite a bit lower, which gives a lower variance. It is also useful to have an optimized negative loglikelihood to compare to other models, as in the development factor example. Here that is 725.

Recall that the mean and variance of each cell differs a little from μ and θμ in the ZMCSP model for the smaller cells. In this case only the last projected column has low values of $\lambda = \mu/\theta$ and these are around 3. This has only a very slight effect on the projected mean and variance. The estimated reserve of 19,334,000 increases by about 1,000 and the standard deviation of 1,230,500 decreases by about 100. Thus in this case that is a very minor impact. Only the change in the estimated θ has any significant influence on the projections.

A good starting point for investigating other possible distributions for the same models structure is fitting the gamma *p*. Aggregate losses are often approximately gamma distributed, and the value of *p* gives an indication of how the variance can be expressed as a multiple of

the mean.

For this data the MLE of p is -0.136, which gives the variance as proportional to the mean raised to 0.864. This is not suggestive of any other popular models, however. The negative loglikelihood is 723.06 compared to 725.00 for the ZMCSP. With 8 parameters compared to 7 for the ZMCSP, the AICc's come out as 732.6 and 733.2, so the gamma p is a little lower. However, if only 6 parameters are counted for the ZMCSP under the view that θ does not affect the fit, its AICc reduces to 731.9. Thus, there is some ambiguity as to which is the best fit. Better ways of counting the degrees of freedom a model uses up would be helpful. In any case the variance is close to proportional to the mean in either model.

Another model with that property is the Poisson-normal. MLE using Mong's formula for $f(x)$ gives m = 35,242 and s = 3,081, with λ's ranging from 2 to 35. The negative loglikelihood is 722.4, which is the best so far. The resulting AICc for 8 parameters is 732.0, which is still ambiguous in comparison to the ZMCSP. The integral for $f(x)$ for the one cell with $\lambda = 2$ is of limited accuracy, so there is a slight degree of additional ambiguity in the value of the AICc.

9 CONCLUSIONS

GLMs provide a powerful modeling tool, but the exponential family has some limitations. By not requiring this form, even familiar distributions can be reparameterized to provide different relationships between the mean and variance of the instances of the fitted dependent variable. When fitting aggregate loss distributions, the gamma is often a good starting point for the shape of the distribution, and so fitting the gamma p, which is a gamma but allows for the variance to be any desired power of the mean, is often a good way to get an indication of the form of the variance to mean relationship. Other distributions can then be tried which have approximately that relationship.

Even when using exponential family distributions, computing power is usually sufficient to calculate the full likelihood function, instead of approximations sometimes used in GLMs. GLZs thus expand the limitations of GLMs, yet there are still situations where it may be useful to use strictly nonlinear models.

REFERENCES

[1] Clark, David R. and Charles A. Thayer. 2004. "A Primer on the Exponential Family of Distributions." *CAS Discussion Paper Program*, 117-148.

[2] Fu, Luyang and Cheng-sheng Peter Wu. 2005. "Generalized Minimum Bias Models." *CAS Forum*,(Winter): 73-121.

[3] Hewitt, Charles C. 1966. "Distribution by Size of Risk-A Model." *PCAS* 53:106-114.

[4] Kaas, Rob. 2005. "Compound Poisson Distributions And GLM's — Tweedie's Distribution." Lecture, Royal Flemish Academy of Belgium for Science and the Arts, http://www.kuleuven.be/ucs/seminars_events/other/files/3afmd/Kaas.PDF.

[5] Klugman, Stuart A., Harry H. Panjer, and Gordon E. Willmot. 2004. *Loss Models: From Data to Decisions*, 2nd Ed. Hoboken, NJ: Wiley.

[6] Mack, Thomas. 2002. *Schadenversicherungsmathematik*, 2nd Ed. Karlsrühe, Ger.:Verlag Versicherungs-wirtshaft.

[7] Mildenhall, Stephen J. 1999. "A Systematic Relationship between Minimum Bias and Generalized Linear Models." *PCAS* 76:393-487.

[8] Mildenhall, Stephen J. 2005. "Discussion of Generalized Minimum Bias Models." *CAS Forum* (Winter): 122-124.

[9] Mong, Shaw. 1980. "Estimating Aggregate Loss Probability and Increased Limit Factor." *CAS Discussion Paper Program:* 358-393.

[10] Renshaw, Arthur E. 1994. "Modeling the Claims Process in the Presence of Covariates." *ASTIN Bulletin* 24, no.:2:265–286.

[11] Taylor, Greg C. and Frank R. Ashe. 1983. "Second Moments of Estimates of Outstanding Claims." *Journal of Econometrics* 23:37-61.

[12] Venter, Gary G. 2007. "Refining Reserve Runoff Ranges." *CAS E-Forum,* Summer (forthcoming)..

Biography of the Author

Gary Venter is managing director at Guy Carpenter, LLC. He has an undergraduate degree in philosophy and mathematics from the University of California and an MS in mathematics from Stanford University. He has previously worked at Fireman's Fund, Prudential Reinsurance, NCCI, Workers Compensation Reinsurance Bureau and Sedgwick Re, some of which still exist in one form or another. At Guy Carpenter, Gary develops risk management and risk modeling methodology for application to insurance companies. He also teaches a graduate course in loss modeling at Columbia University.

917.937.3277 **gary.g.venter@guycarp.com**