



# Alternative to the Tweedie in Pure Premium GLM

RPM Seminar 2022

March 15, 2022

David R. Clark

Munich Re America – Corporate Risk and Underwriting



# Agenda

1. Why Tweedie?
2. Collective Risk Model interpretation
3. Variance Structures
4. Extra: Estimating Equations

# Why Tweedie?

Introduced in the insurance context in 1994 by Jorgensen et al, “Fitting Tweedie’s compound Poisson model to insurance claims data.”

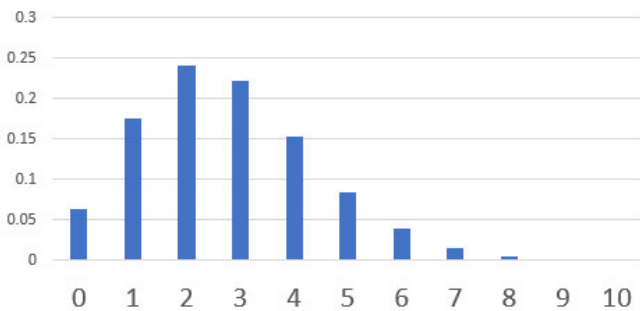
It appears to have been adopted in the US for use in pure premium GLM ratemaking, without considering alternatives.

The attraction of the Tweedie distribution is two-fold:

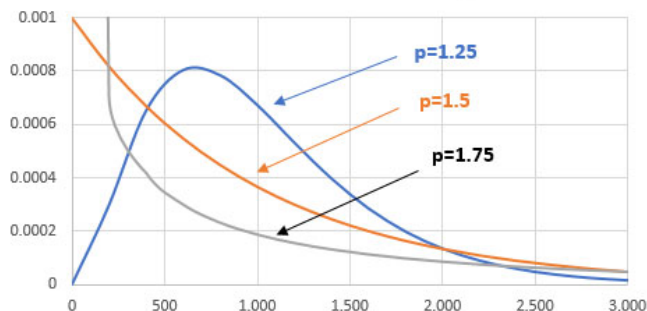
- It can be interpreted as a collective risk model (i.e., a combination of frequency and severity)
- It includes a mass point as zero – for “unbalanced” data

# Tweedie as Collective Risk

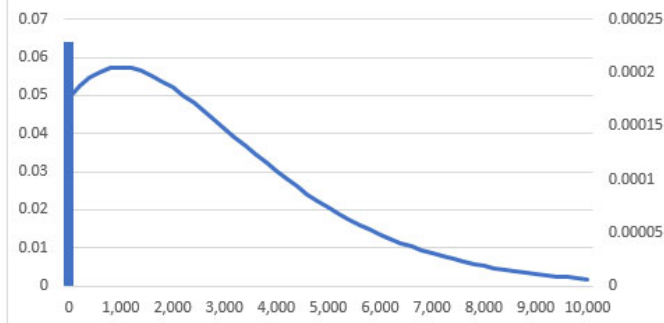
### Poisson Frequency



### Gamma Severity



### Aggregate Distribution



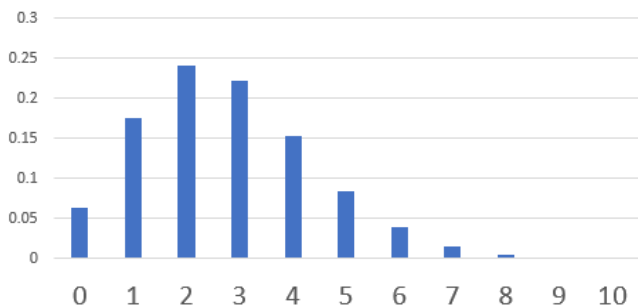
The Tweedie distribution can be interpreted as a collective risk model with Poisson frequency and gamma severity.

A variance parameter “ $p$ ” determines the shape of the gamma severity, with  $p=1.5$  representing an exponential distribution.

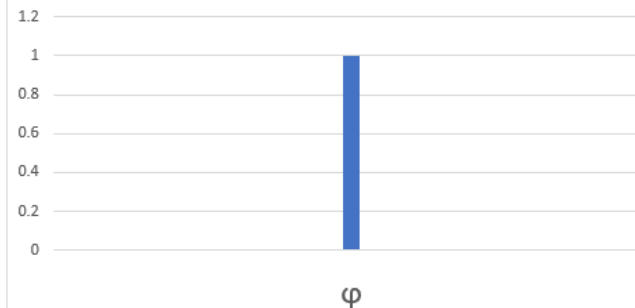
The aggregate distribution is of mixed type, with a mass point at zero and continuous curve for larger values.

# Over-dispersed Poisson (ODP) as Collective Risk

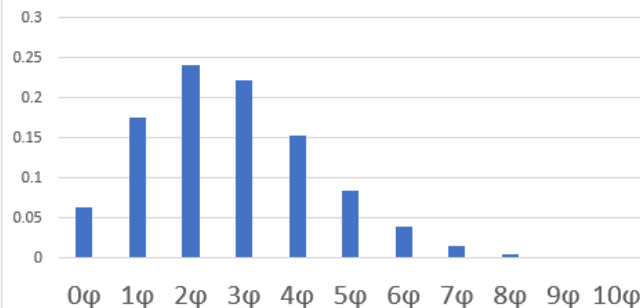
Poisson Frequency



Constant Severity



Aggregate Distribution



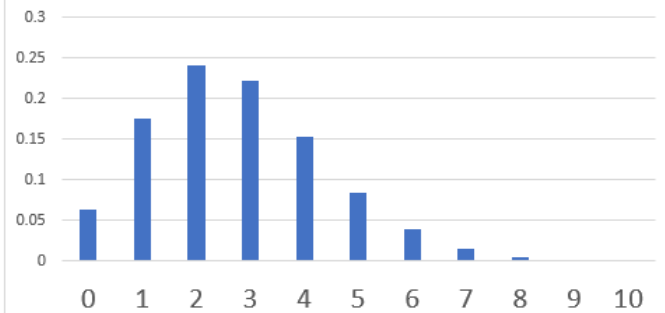
The over-dispersed Poisson (ODP) is the usual name for GLM when variance is assumed to be proportional to expected value.

We never explicitly use this as a distribution form, but it can be interpreted as a simple collective risk model with Poisson frequency and constant severity (Gary Venter likes to call it PCS instead of ODP).

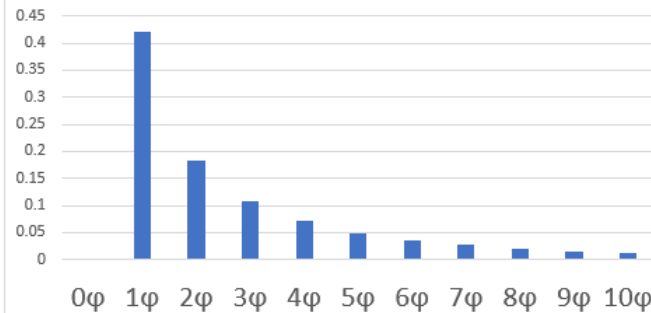
The aggregate distribution is discrete and equal to Poisson times a scale factor.

# Quasi Negative Binomial (QNB) as Collective Risk

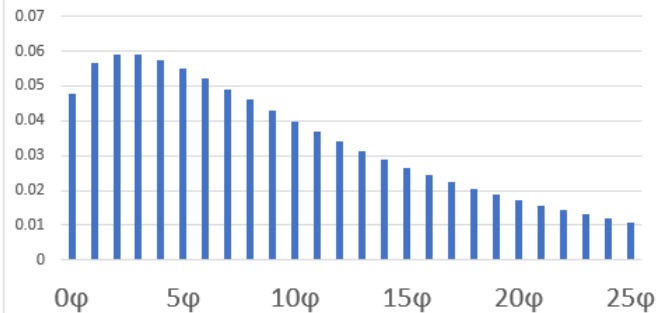
### Poisson Frequency



### Logarithmic Severity



### Aggregate Distribution



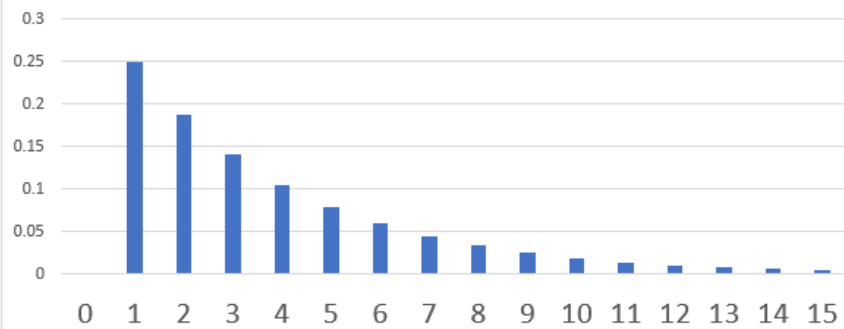
Proposed Alternative to Tweedie:

The quasi negative binomial can also be interpreted as a collective risk model with Poisson frequency and logarithmic severity.

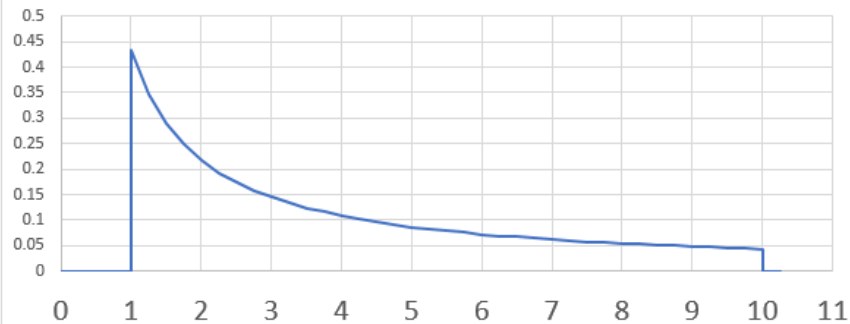
The logarithmic severity is thicker tailed than the Gamma distribution.

# Logarithmic Severity ?!?

### Geometric Distribution



### Upper-Truncated Pareto



The logarithmic severity starts with a geometric distribution.

Geometric is discrete with each probability a constant ratio to the probability for the lower value. For QNB we also shift it to start at 1 rather than 0.

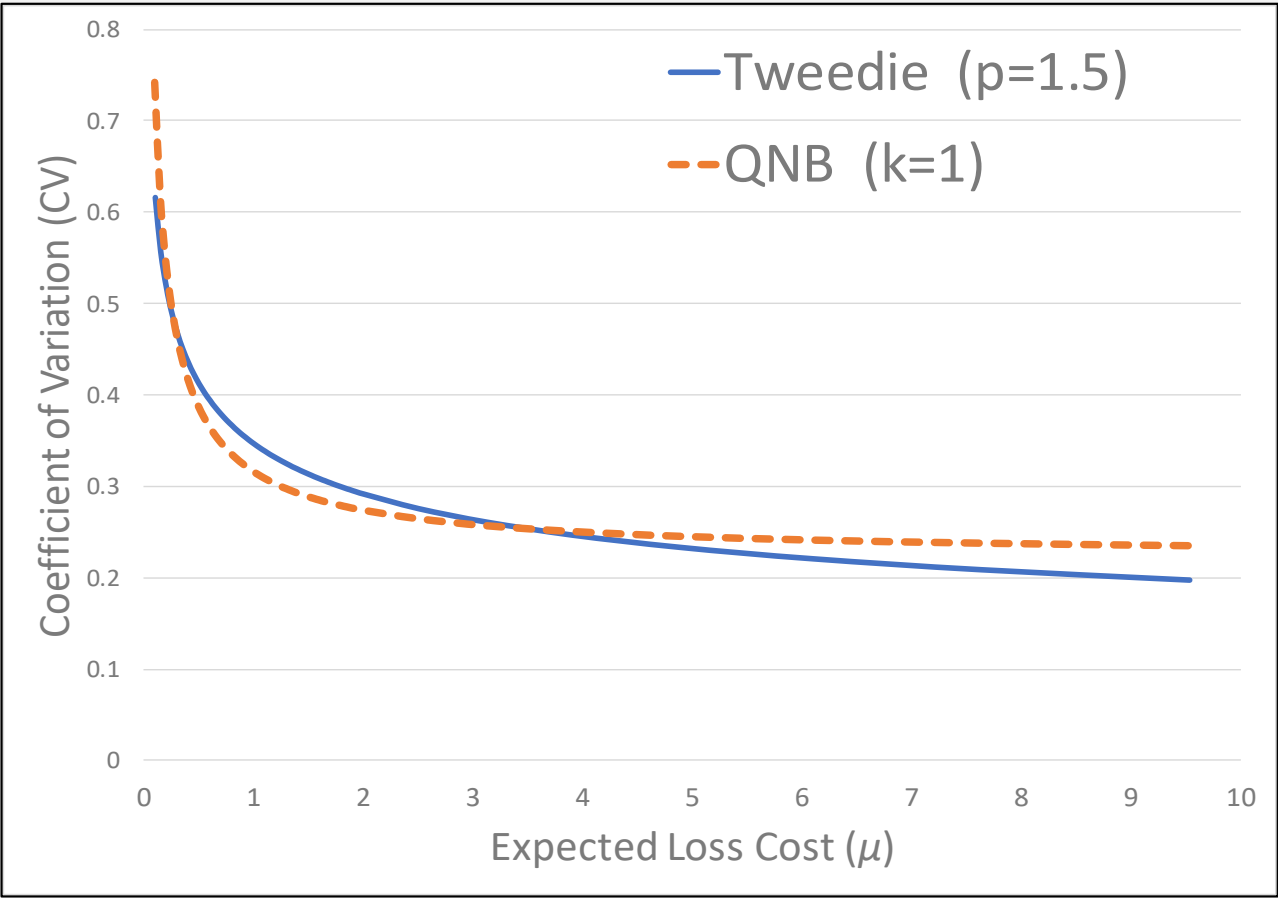
Geometric is the discrete analogy to an exponential distribution.

The mean of the geometric distribution is “mixed” using a form of upper truncated pareto.

$$f(\mu) = \frac{1}{\ln(M)} \cdot \frac{1}{\mu} \quad 1 \leq \mu \leq M$$

The logarithmic distribution can be considered a discrete version of a mixed exponential model.

# Comparison of Tail Behavior



The QNB is a modestly thicker tailed distribution than the Tweedie.

The QNB will be slightly less sensitive to extreme values of very low or very high pure premium values.

See Appendix for more details on the higher moments.



## Variance Structure: Tweedie

For GLM, we do not need the full collective risk model interpretation.

The conditions are relaxed such that we only need to know the relationship of the variance as a function of the mean.

For Tweedie:

$$V(\mu_i) = \phi \cdot \mu_i^p \quad 1 \leq p \leq 2$$

The dispersion parameter  $\phi$  is considered a “nuisance parameter” that does not affect the expected pure premium fit.

The variance parameter  $p$  is supplied by the model user.

Special cases:

$p = 1$  is over-dispersed Poisson

$p = 2$  is Gamma

## Variance Structure: Quasi Negative Binomial

For Quasi Negative Binomial (QNB):

$$V(\mu_i) = \phi \cdot \left( \mu_i + \frac{1}{k} \cdot \mu_i^2 \right)$$

The dispersion parameter  $\phi$  is considered a “nuisance parameter” that does not affect the expected pure premium fit (same as in Tweedie).

The variance parameter  $k$  is supplied by the model user.

As with the Tweedie, the QNB is a compromise between ODP and Gamma variance structures.

[QNB is an arithmetic (additive) combination, Tweedie is a geometric (multiplicative) combination]

## Variance Structure: Collective Risk Models

In usual notation, aggregate losses  $Z$  are treated as a random sum of severity  $X$  and frequency  $N$ .

$$Z = \sum_{j=1}^N X_j$$

$$\text{Var}(Z) = \text{Var}(X) \cdot E(N) + E(X)^2 \cdot \text{Var}(N)$$

If the frequency is Negative Binomial, with a contagion  $c$ , then the variance formula can be re-written as:

$$\text{Var}(N) = E(N) + c \cdot E(N)^2$$

$$\text{Var}(Z) = \left( \frac{E(X^2)}{E(X)} \right) \cdot E(Z) + c \cdot E(Z)^2$$

This form is equivalent to the QNB variance structure.

*“Anything you can do I can do better”*

Irving Berlin (Annie Get Your Gun)

Why Quasi Negative Binomial (QNB) instead of Tweedie?

- Just as easy to implement in GLM framework
- QNB can also be interpreted as a collective risk model
- It is a compromise between Poisson and Gamma variance structures
- Variance parameter is easier to compare to other collective risk models
- Slightly thicker tail behavior

### *“I Don’t Want to Change the World”*

Ozzy Osbourne

In practice, the choice of variance function will not materially change the fitted values in a classification rating plan.

Getting the variance function wrong is a form of heteroskedasticity. This may distort the significance statistics (e.g., t-statistics) and even change the decision on which variables to include.

But if all you need is to fit the expected values, then the choice of Tweedie versus QNB is not critical.

*“There is so much more I could have done if they'd let me!”*

Nick Cave (The Curse of Millhaven)

The CAS call for essay constrained the paper to three pages, so limited the detail that could be included.

Variance structures in GLM can also be compared by looking at their Estimating Equations.

This gives more insight into how GLM makes use of the variance assumptions.

# Estimating Equations

As noted above, the different variance structures can be interpreted as collective risk models.

But in GLM, they do not have to be!

Rather than thinking in terms of distributions, GLM more naturally works with weighted averages.

Estimating Equations could better be thought of as balance equations:

*Under what weighting scheme do the fitted values balance to the actual values?*

For the ODP, with the “canonical” log-link (rating factors applied multiplicatively), the fitted values balance to the actual values for any column of predictors.

$$\sum_{i=1}^n x_{i,j} \cdot y_i = \sum_{i=1}^n x_{i,j} \cdot \mu_i \quad \forall j$$

# Estimating Equations

We can also include weights  $w_i$  to give added flexibility to the balancing equations.

$$\sum_{i=1}^n x_{i,j} \cdot w_i \cdot y_i = \sum_{i=1}^n x_{i,j} \cdot w_i \cdot \mu_i$$

The result in classification ratemaking is that the fitted pure premium values balance to the actual losses across each dimension of the rating plan.

In loss development work, this is why the ODP GLM matches the chain ladder method.

If we stay with the log-link structure but switch to Tweedie or QNB variances, the weights are adjusted by a function of the fitted values.



## Estimating Equations – Tweedie with log-link

We keep the log-link structure (rating variables applied multiplicatively) but change to Tweedie variance.

$$\sum_{i=1}^n x_{i,j} \cdot (\mu_i^{1-p}) \cdot y_i = \sum_{i=1}^n x_{i,j} \cdot (\mu_i^{1-p}) \cdot \mu_i$$

The estimating equations show the weighting scheme under which the fitted values match the actual values. Admittedly, there is not much intuition to these weights.

Note also: the idea of the “mass point at zero” does not play any role in this.

## Estimating Equations – QNB with log-link

We keep the log-link structure (rating variables applied multiplicatively) but change to QNB variance.

$$\sum_{i=1}^n x_{i,j} \cdot \left( \frac{k}{\mu_i + k} \right) \cdot y_i = \sum_{i=1}^n x_{i,j} \cdot \left( \frac{k}{\mu_i + k} \right) \cdot \mu_i$$

The weights (in parenthesis) are constrained between 0 and 1, so are a bit more stable than for the Tweedie.

We can also re-arrange the terms as below. The right-hand side looks like an experience-rating formula.

$$\sum_{i=1}^n x_{i,j} \cdot y_i = \sum_{i=1}^n x_{i,j} \cdot \left[ \left( \frac{\mu_i}{\mu_i + k} \right) \cdot y_i + \left( \frac{k}{\mu_i + k} \right) \cdot \mu_i \right]$$

# References

Andersen, Duncan, Sholom Feldblum, Claudine Modlin, Doris Schirmacher, Ernesto Schirmacher, and Neeza Thandi, “A Practitioner’s Guide to Generalized Linear Models,” CAS Study Note, 2007.

Clark, DR. “A Note on the Upper-Truncated Pareto Distribution,” CAS Forum Winter 2013, Vol. 1, 1-22.

Goldburd, Mark, Anand Khare, Dan Tevet, and Dmitriy Guller, “Generalized Linear Models for Insurance Ratemaking,” CAS Monograph No. 5.

Hilbe, Joseph M., “Negative Binomial Regression,” Second Edition, 2011, Cambridge University Press,

Jorgenson B, de Souza MCP, “Fitting Tweedie’s compound Poisson model to insurance claims data,” Scandinavian Actuarial Journal, 1994. 69-93.

R. W. M. Wedderburn, “Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method,” Biometrika 1974, Vol. 61 No. 3, 439-447.

## Appendix: Comparison of Higher Moments

Just for reference, we can also confirm that the higher moment for kurtosis is slightly higher for the QNB.

	<u>Tweedie</u>	<u>Quasi-Negative Binomial</u>
Mean	$\mu$	$\mu$
Variance	$\phi \cdot \mu^p$	$\phi \cdot \left( \mu + \frac{1}{k} \cdot \mu^2 \right)$
Coefficient of Variation	$CV = \sqrt{\frac{\phi}{\mu^{2-p}}}$	$CV = \sqrt{\frac{\phi}{\mu} + \frac{\phi}{k}}$
$\lim_{\mu \rightarrow \infty} CV$	0	$\sqrt{\frac{\phi}{k}}$
Skewness	$p \cdot CV$	$\left( 1 + \frac{\mu}{\mu + k} \right) \cdot CV$
Kurtosis	$3 + p \cdot CV \cdot (Skew - CV) + Skew^2$	$3 + 2 \cdot CV \cdot (Skew - CV) + Skew^2$

## Appendix: Variance Functions and Quasi-Likelihood

Wedderburn (1974) introduced the quasi-likelihood function as a way to extend GLM beyond explicit distribution forms. It is defined using only the variance function.

$$\text{Quasi-Likelihood} = \int_y^\mu \frac{y-t}{\phi \cdot V(t)} dt$$

The variance functions for the models we have discussed are given below.

(overdispersed) Poisson	$V(\mu) = \mu$
Tweedie	$V(\mu) = \mu^p$
Quasi Negative Binomial	$V(\mu) = \mu + \frac{1}{k} \cdot \mu^2$

# Imprint

## Munich Reinsurance

MRAS – Corporate Risk and Underwriting Services

David R. Clark

[dclark@munichre.com](mailto:dclark@munichre.com)

© 2022 Münchener Rückversicherungs-Gesellschaft

© 2022 Munich Reinsurance Company

# Alternative to the Tweedie in Pure Premium GLM - Comparing Tweedie, Quasi-Poisson, and Quasi-Negative Binomial

---

RPM 2022

Josh Brady, FCAS, MAAA, PhD

The Cincinnati Insurance Companies

# Background

---

## Model Framework

- Loss cost GLM with log link
- Examples use public dataset datacar (see appendix)

## What is quasi-Poisson?

- Like Poisson except defined on all non-negative values instead of just integers
- Variance relationship allows for over-dispersion:  $Var(y) = \phi\mu$

## Quasi-Poisson connection to estimating equations/minimum bias are well known

- Mildenhall: A SYSTEMATIC RELATIONSHIP BETWEEN MINIMUM BIAS AND GENERALIZED LINEAR MODELS

## More comprehensive presentation on quasi-Poisson

- Presented at BACE regional affiliate meeting in 2019
- <https://community.casact.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=394ac193-30d9-b9f9-a6f1-b5ddd03da6a3&forceDialog=0>



# Distribution Comparisons

---

## Quasi-Poisson (QP)

- Predictions are balanced to observations on categorical co-variates
- Simplifies the offset process (exposure offset is equivalent to model offset)
- Testing in R and SAS shows quasi-Poisson models fit faster
  - Human time in model building is a cost/constraint

## Tweedie

- Tweedie appears to be a more appropriate distribution as compared to quasi-Poisson when tested on loss cost data
- Tweedie has better predictive power as compared to quasi-Poisson
- Tweedie is more common/accepted

## Quasi-Negative Binomial (QNB)

- Less influenced by extreme observations
- Potentially more stable convergence
- Collective risk model interpretation

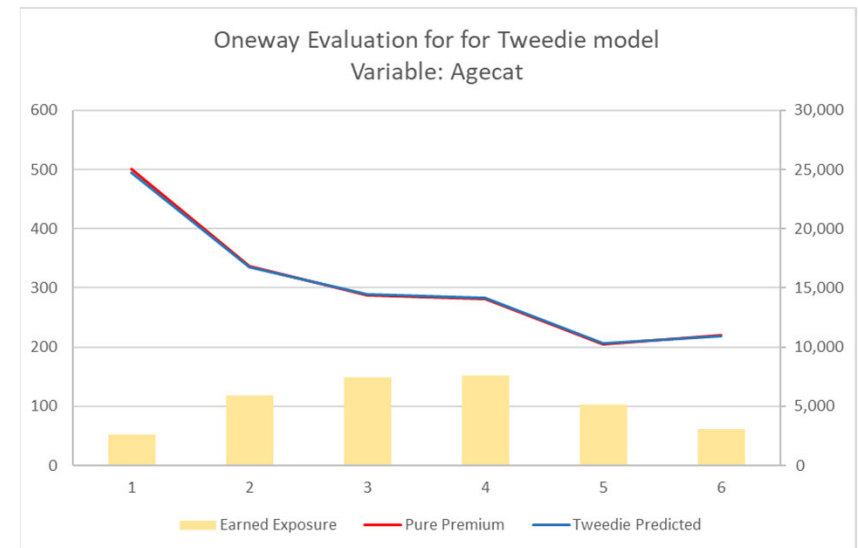
For Tweedie vs QP, my experience is that the observations above hold broadly  
I haven't performed extensive testing for Tweedie vs QNB beyond the examples in this presentation

# Initial Motivations

– Tweedie predictions are not balanced to observations

- Fitted GLM on sample dataset (datacar) using a Tweedie distribution ( $p=1.5$ )
- Included categorical covariate agecat in the model
- We can see for Tweedie that the predicted pure premium  $\neq$  actual pure premium
- Using the quasi-Poisson distribution, the predictions are in fact balanced to the observed pure premium

agecat	Earned Exposure	Pure Premium	Tweedie Predicted	Tweedie %Difference	QP Predicted
1	2,612	500	495	-1.2%	500
2	5,892	337	335	-0.5%	337
3	7,409	288	289	0.3%	288
4	7,617	282	283	0.6%	282
5	5,171	205	206	0.5%	205
6	3,100	221	219	-0.6%	221
Total	3,100	293	293	0.0%	293



The mismatches for Tweedie occur even on large datasets with credible data. The cause is due to the model specification.

# Why does Tweedie have bias in the predictions?

## Answer: Variance structure and link function

---

- When the GLM link function is the canonical link for the distribution then the predictions will be balanced to the observations
- We almost always use a log link, which is not the canonical link for the Tweedie distribution
  - Log link is the canonical link for Poisson
- Balance equations termed in loss vs predicted loss
  - Tweedie parameter  $p$
  - Notice that when  $p = 1$  (quasi-Poisson) then predicted losses balance to actual losses
  - When  $p \neq 1$  then in general predictions will not balance

$$\sum_i \text{loss}_i \mu_i^{1-p} X_{ij} = \sum_i [\text{pred loss}]_i \mu_i^{1-p} X_{ij}$$

# What's wrong with regular Poisson? Why do we need quasi-Poisson?

---

- There is a fundamental issue to using the Poisson distribution for loss cost modeling
  - The Poisson distribution is not defined for non-integers
  - Poisson probability:

$$f(y|\mu) = \frac{e^{-\mu} \mu^y}{y!}$$

- A second issue with the Poisson distribution is that it is assumed that the variance equals the mean
  - $Var(y) = \mu$
- Quasi-Poisson improves this relationship by allowing the variance to be proportional to the mean:
  - $Var(y) = \phi\mu$

# Can we extend the Poisson distribution to all non-negative values?

---

Answer: No

- Pmf for the Poisson distribution with mean  $\mu$ :
  - $f(k|\mu) = e^{-\mu} \frac{\mu^k}{k!}, k \in \{0,1,2, \dots\}$
- Suppose we wanted to extend the Poisson distribution from integers to all non-negative numbers in a way such that the parameter estimates were unchanged
- That is, can we replace  $k!$  with a (reasonably nice) function  $g$  such that for  $y > 0$

$$f_2(y|\mu) = e^{-\mu} \frac{\mu^y}{g(y)}$$

is a probability distribution?

- Natural candidate would be  $g(y) = \Gamma(y + 1)$
- Turns out that it is not possible
  - Why?
  - For the curious... Proof relies on showing the moment generating functions are equal on an open domain. Result is to conclude distributions are in fact the same.

# Sample dataset performance comparisons

---

- Examples are great, but we must be careful in generalizing from these examples
- Tweedie vs quasi-Poisson: I have found similar results across many different datasets
- I haven't performed much testing with quasi-negative binomial beyond these examples

# Quasi-Negative Binomial Testing Comments

---

Variance relationship:  $Var(y) = \phi(\mu + \frac{1}{k}\mu^2)$

- The parameter  $k$  determines the mixing between the linear term and the quadratic term
- A difficulty with building intuition is that the  $k$  is not unitless and needs to be compared to the mean for intuition
- The Tweedie  $p$  is unitless and selecting a value around 1.6-1.8 is often a reasonable starting point

## Testing QNB in R

- `glm.nb` allows estimation of both the variable coefficients and  $k$  (MLE estimation)
- On the test dataset `glm.nb` found that  $k \approx 0.01$
- We can also use R `glm` with the negative binomial distribution
  - Need to specify  $k$
  - Didn't converge on the sample dataset for  $k \leq 2$
- Both `glm.nb` and `glm` do not currently utilize quasi-frameworks. That is, the distributions do not allow for overdispersion. The functions do accommodate non-integer observations.
  - $Var(y) = \mu + \frac{1}{k}\mu^2$
  - It is not difficult to calculate a Pearson or deviance estimate of  $\phi$ . Then this estimated dispersion parameter can be used to calculate significance tests that contemplate overdispersion.

# Performance - Gini

---

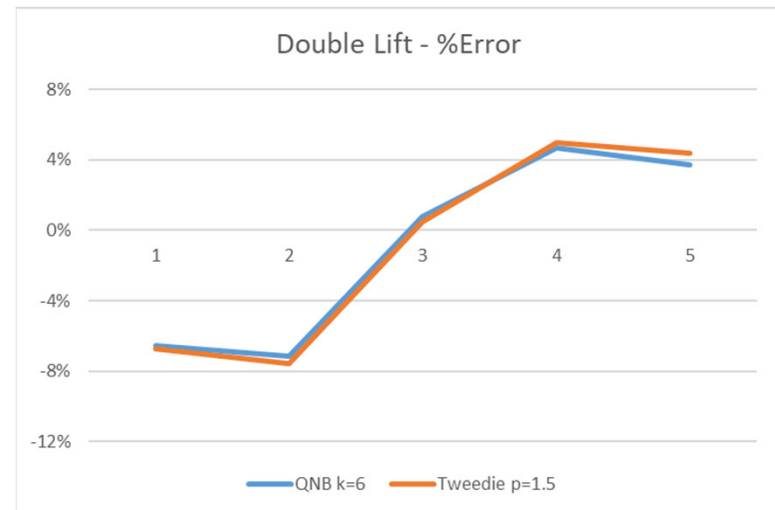
- Fit model and evaluate on entire dataset (in-sample testing)
- Cross-Val (10 fold, 5 times)
- Note that this is one small sample data set
  - Requires 4-5 decimal places to see differences
- Results
  - Tweedie has the best cross validated performance
  - Negative binomial (NB) between Tweedie and quasi-Poisson
  - Quasi-Poisson (QP) shows the smallest drop in performance

Model	Full Dataset	Cross-Val	Difference
QP	0.14114	0.12367	0.01748
Tweedie ( $p = 1.2$ )	0.14130	0.12374	0.01755
Tweedie ( $p = 1.5$ )	0.14173	0.12374	0.01799
Tweedie ( $p = 1.7$ )	0.14205	0.12374	0.01831
NB ( $k = 3$ )	0.14233	0.12370	0.01864
NB ( $k = 4$ )	0.14230	0.12370	0.01860
NB ( $k = 5$ )	0.14228	0.12371	0.01857
NB ( $k = 6$ )	0.14230	0.12372	0.01859













# Performance – Double Lift

- 5 bin cross validated double lift
- Charts show %error = (pred – act)/act
- Left chart shows that Tweedie has lower error vs quasi-Poisson
- Right chart shows that quasi-negative binomial has lower error than Tweedie
- Once again, this is just one sample data set



# Algorithm Speed – R glm

Model	Relative Time	
QP	1.00	
NB (k estimated)	4.06	
NB (k=3)	1.47	
NB (k=4)	1.07	
NB (k=5)	0.87	
NB (k=6)	0.75	
Tweedie (p=1)	2.09	
Tweedie (p=1.2)	1.61	
Tweedie (p=1.5)	1.25	
Tweedie (p=1.7)	2.37	

- Overall Tweedie appears to be at least 30% slower than quasi-Poisson on the test data set
- QNB can be faster than QP for a fixed k, the issue is that k needs to be estimated which greatly increases fitting time

NB = Negative binomial (recall that R doesn't have a quasi-framework yet)

Dispersion estimate

- R uses the Pearson estimator for dispersion
  - $r$  is the number of parameters

$$\phi = \frac{1}{n-r} \sum_i \frac{(y_i - \mu_i)^2}{v(\mu_i)}$$








Tweedie likelihood

- Likelihood is not computed as it is computationally expensive (can use tweedie package to compute)
- I didn't compare times for estimating the Tweedie  $p$  (can use tweedie.profile)

# Algorithm Speed – SAS hpgenselect

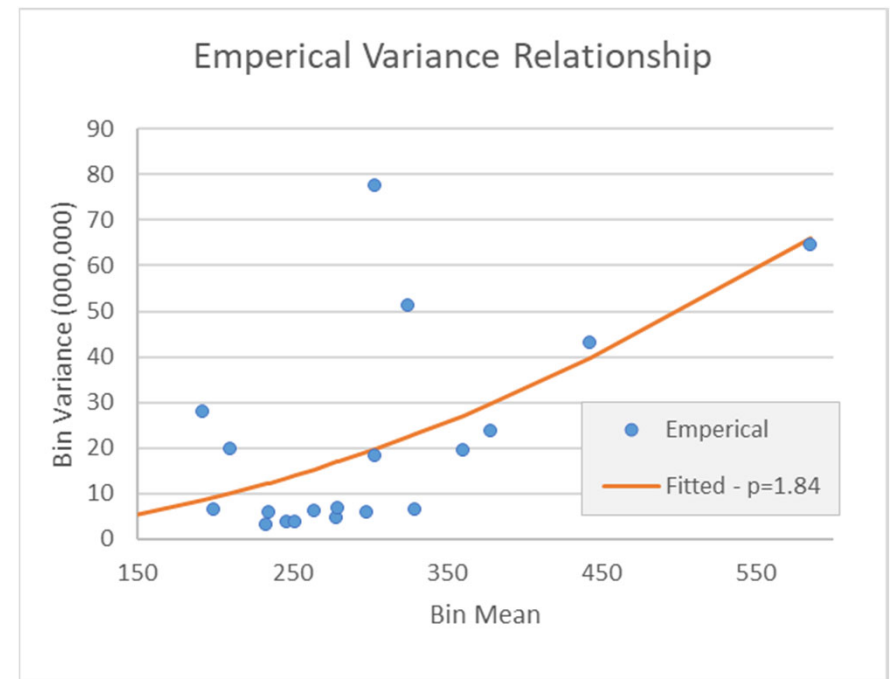
---

- SAS hpgenselect has the option to use maximum likelihood to estimate the dispersion  $\phi$  and power  $p$  for the Tweedie distribution
  - Tweedie ( $p$  MLE estimated) fits both  $p$  and  $\phi$  using MLE
  - Tweedie (1.5) –  $p$  was specified to be 1.5 and  $\phi$  is still estimated using MLE
  - quasi-Tweedie is an option to avoid MLE estimation of  $\phi$  (similar to R)
- For Poisson and negative binomial hpgenselect doesn't estimate an overdispersion parameter  $\phi$ , though the Pearson estimate is simple to compute
- Poisson is fastest, whereas Tweedie is the slowest even when using the quasi-Tweedie

<b>Model</b>	<b>Relative Time</b>	
Poisson	1.00	
Tweedie (p MLE estimated)	18.32	
Tweedie (p=1.5)	5.78	
quasi-Tweedie wDispersion Est (p=1.5)	5.53	
quasi-Tweedie woDispersion Est (p=1.5)	3.66	
NB (k estimated)	2.29	
NB (k = 0.01)	1.28	

# Examining the variance relationship of the sample data

- Fit quasi-Poisson model for predicted mean
  - Wanted to give QP the best shot at showing a linear relationship. Using Tweedie predictions give similar results.
- Ranked low to high and binned into 20 equal exposure bins
- On each bin calculated empirical mean and variance
- Plot to right shows the relationship
- Fit curve of form  $\text{Var}_{\text{bin}} = a \mu_{\text{bin}}^p$ 
  - a is the intercept
  - p is the fitted power (1.84 in this case)
- Suggests that a Tweedie variance relationship is more appropriate than quasi-Poisson (linear)
  - Also implies that the QNB variance structure may be more appropriate



# Factor Offsets

---

- Suppose in our model we wanted to apply fixed factors
  - These could be in the current model or perhaps previously selected factors
- With loss cost data two of the most common ways to apply factors offsets are:
  - Let  $F$  denote multiplicative factors to offset
    1. Model offset:  $\frac{\text{loss}}{EE} \sim \eta + \log(F)$ , weight =  $EE$
    2. Exposure offset:  $\frac{\text{loss}}{EE * F} \sim \eta$ , weight =  $EE * F$
- For the QNB and Tweedie the exposure offset is not equivalent to the model offset, whereas for QP the two methods of offsetting are equivalent (Shi 2010)
- Why does this work for QP?
  - Answer: The log link is the canonical link for the Poisson distribution

# Conclusion

---

Any of these distributions can be a reasonable choice for modeling loss cost

## ❖ Quasi-Poisson

- Fits faster
- Predictions are balanced to losses for categorical variables
- Exposure offset is equivalent to model offset

## ❖ Tweedie

- Variance structure appears more appropriate as compared to quasi-Poisson
- Best cross validated performance on sample dataset

## ❖ QNB

- Heavier tailed → extreme observations have less influence
- Collective risk model interpretation
- Double lift was superior on sample data set
- Similar (perhaps faster) fitting time to QP
- R implementation issues: lack of convergence for small  $k$  on our sample dataset does not currently consider overdispersion (can be handled manually)
- Issue is that the parameter  $k$  needs to be estimated. Currently there are heuristics that allow for a reasonable default value.
  - E.g.,  $p = 1.6$  to  $1.8$  for Tweedie is accepted by many as reasonable for initial model building

# Contact Information

---

[Joshua\\_Brady@cifin.com](mailto:Joshua_Brady@cifin.com)

<https://www.linkedin.com/in/joshua-brady-fcas/>

# References

Bailey, Robert A., and LeRoy J. Simon. "Two studies in automobile insurance ratemaking." *ASTIN Bulletin: The Journal of the IAA* 1.4 (1960): 192-217.

Dunn, Peter K., and Gordon K. Smyth. "Evaluation of Tweedie exponential dispersion model densities by Fourier inversion." *Statistics and Computing* 18.1 (2008): 73-86.

Jorgensen, Bent. *The theory of dispersion models*. CRC Press, 1997.

McCullagh, Peter. *Generalized linear models*. Routledge, 2018.

Mildenhall, Stephen J. "A systematic relationship between minimum bias and generalized linear models." *Proceedings of the Casualty Actuarial Society*. Vol. 86. No. 164. 1999.

Ohlsson, Esbjörn, and Björn Johansson. *Non-life insurance pricing with generalized linear models*. Vol. 2. Berlin: Springer, 2010.

Shi, Sheng G. "Direct analysis of pre-adjusted loss cost, frequency or severity in tweedie models." *Casualty Actuarial Society E-Forum*. 2010.



# Appendix

# Data set used to contrast model performance

---

## **dataCar\***

- This data set is based on one-year vehicle insurance policies taken out in 2004 or 2005.
- 67856 observations
- Frequency ~ 15.5%
- Severity ~ \$1900

### Fields used in sample model

- claimcst0 - loss
- Exposure
- Pure premium (pp) = claimcst0/exposure
- veh\_value in \$10,000s
- veh\_body
  - Categorical with levels BUS CONVT COUPE HBACK HDTOP MCARA MIBUS PANVN RDSTR SEDAN STNWG TRUCK UTE
- gender
  - categorical with levels F M
- agecat
  - 1 (youngest), 2, 3, 4, 5, 6

## **Model**

### Data Adjustments

- veh\_body\_grp2 = grouped small exposure levels with other levels
- veh\_val5 = vehicle value rounded to nearest 0.1 and capped at 5
- Transform agecat to factor to model as categorical variable

Target = Pure Premium (pp) = Claimcst0/exposure

Weight = exposure (EE)

Formula:

$pp \sim agecat + gender + veh\_body\_gp2 + veh\_val5$

\*Contained in insuranceData R package <https://cran.r-project.org/web/packages/insuranceData/index.html>