Journal of Advances in Mathematics and Computer Science

30(1): 1-13, 2019; Article no.JAMCS.45962
ISSN: 2456-9968
(Past name: British Journal of Mathematics & Computer Science, Past ISSN: 2231-0851)

Modeling with the Power Variance form of the Gamma Distribution

Katherine E. Irimata[1*], N. David Yanez[2], Ibrahim A. Aljasser[3] and Jeffrey R. Wilson[4]

[1]Division of Research and Methodology, National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, USA.
[2]School of Public Health, Oregon Health & Science University, Portland, Oregon, USA.
[3]Department of Quantitative Analysis, King Saud University, Riyadh, Saudi Arabia.
[4]Department of Economics, Arizona State University, Tempe, Arizona, USA.

Authors' contributions

This work was carried out in collaboration between all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JAMCS/2019/45962
Editor(s):
(1) Dr. Kai-Long Hsiao, Associate Professor, Taiwan Shoufu University, Taiwan.
Reviewers:
(1) Bachir Achour, University of Biskra, Algeria.
(2) Zlatin Zlatev, Trakia University, Bulgaria.
(3) Suchandan Kayal, NIT Rourkela, India.
(4) Ahmed F. I. Siddiqi, UCP Business School, University of Central Punjab, Pakistan.
Complete Peer review History: http://www.sciencedomain.org/review-history/28042

Original Research Article

Received: 01 October 2018
Accepted: 12 December 2018
Published: 01 January 2019

Abstract

It is not uncommon to encounter data where the distribution of the responses is not known to completely follow any of the common probability models. While there are general classes of models, such as the Tweedie distribution, which can be adopted in such cases, many approximations have been proposed based on the fact that they are often easier to obtain. We bring to the discussion a three-parameter power variance representation of the gamma distribution $\Gamma(\alpha, \beta)$ that has a general mean-variance relationship $Var(Y) = \phi\mu^{\tau}$, where $\mu = E(Y)$ is the mean or expected value of Y, ϕ is a scale parameter, and τ is the degree of power of the expression. This power variance formulation is a flexible extension of the gamma distribution, and are used to approximate various models and determine significant predictors even when the distribution is not fully realized. We present a comparison of the power variance model to several known distributions which have similar mean-variance. In addition, we provide a more general representation of the relation $Var(Y) = \phi V_{\tau}(\mu)$, where $V_{\tau}(\mu)$ is the variance function indexed by the

**Corresponding author: E-mail: kirimata@cdc.gov, katherine.irimata@gmail.com;*

parameter τ. We demonstrate the performance of the power variance modeling approach through a simulation and evaluate two numerical examples, including high school absenteeism and concrete compression strength.

Keywords: Positive random variables; three-parameter gamma; mean-variance relationship; Tweedie distribution.

1 Introduction

It is common to begin the analysis of data by assuming that the responses follow an assumed distribution. This is usually influenced by certain defining features. For example, it is well known that the evaluation of count data begins with the assumption of the responses following a Poisson distribution. This may include the number of days until a patient is released from the hospital or the number of complications a patient experiences post-surgery. Although these types of data are commonly analyzed using Poisson regression with the count as the outcome, it may be the case that the data do not have follow our assumption of the standard Poisson distribution [1, 2]. This may be due to a certain extraneous factors, including the presence of overdispersion based on the survey design or the fact that the data were obtained based on some hierarchical structure. More so, it may be the case that the distribution selected to describe the responses in the data lacks the appropriate shape due to deviations in skewness or kurtosis. For example, although we often select the Poisson distribution in the evaluation of count data, the fit is affected when dealing with count data with an excess of zeroes.

To characterize data, one often describes the distribution by relying on a description of the first two moments, the mean $\boldsymbol{\mu}$ and variance $Var(\boldsymbol{Y})$. As in the exponential family of distributions, the variance is often related to some function of the mean, and such is the parameterization implemented in the Tweedie distribution. The Tweedie distribution [3], a class of exponential dispersion models, is characterized by the parameters ϕ, $\boldsymbol{\mu}$, and τ such that a mean-variance relationship is $Var(\boldsymbol{Y}) = \phi V(\boldsymbol{\mu}) = \phi\mu^{\tau}$, where $E(\boldsymbol{Y}) = \boldsymbol{\mu}$. The Tweedie probability density function is not available in closed form, except in special cases, and then requires complex numerical approximations to evaluate the density. In particular, for the case where $1 < \tau < 2$, the density function is written as a function of θ, ϕ, and α where μ is a function of θ and $\tau = \frac{\alpha-2}{\alpha-1}$ such that

$$f_Y(y;\theta,\phi,\alpha) = \sum_{n=1}^{\infty} \frac{\left[\left(\frac{\omega}{\phi}\right)^{1-\alpha}\kappa_\alpha\left(-\frac{1}{y}\right)\right]^n}{\Gamma(-n\alpha)n!y} exp\left\{\frac{\omega}{\phi}[\theta_0 y - \kappa_\alpha(\theta_0)]\right\} \text{ for } y > 0$$

Where

$$\kappa_\alpha(\theta) = \frac{\alpha-1}{\alpha}\left(\frac{\theta}{\alpha-1}\right)^{\alpha}, \theta_0 = \theta\phi^{1-\alpha},$$

and ω is the prior weight [4]. The special cases of the Tweedie distribution include the normal (τ = 0), Poisson (τ = 0 and ϕ = 0), gamma (τ = 2), and inverse Gaussian (τ = 3). Thus, for these density functions (τ = 0, 1, 2, 3) one can directly compute the Tweedie distribution. Alternative estimation methods include series expansions, maximum likelihood estimation, and quasi-likelihood estimation [5-7]. Also, there is software available to evaluate the Tweedie distribution for $\tau \geq 1$ [8].

It is often necessary in the data analysis to have at least some partial distributional assumption for the response even when the probability model is not completely realized. In those case one can rely on the Tweedie distribution to describe the mean and variance of the data, if one is comfortable with estimation under the special cases. However, making use of general forms of alternative distributions has shown to produce good approximations and are often easier to estimate. Stacy [9] proposed a three-parameter gamma distribution which includes many distributions as special cases. Stacy and Mihram [10] examined the

generalization of the gamma distribution and derived a parameter estimation technique using a modified method of moments approach with a graphical aid. They demonstrated that enlarging the group of probability distributions and considering alternative estimation methods improve the accuracy of numerical analyses. Wilks [11] investigated various three-parameter probability distributions and found that they provided additional flexibility and improved probability predictions compared to traditional distributions. These studies demonstrate the tractability of the generalized gamma distribution. We focus on the gamma distribution as it has many different parameterizations and its flexibility allows it to be used in the approximation of many distributions and in modeling unknown cases.

As such, we present the power variance (PV) distribution, a three-parameter gamma distribution parameterized using the mean and variance relationship with parameters ϕ, τ and μ. This power variance form provides an extension to the gamma family, but has an additional flexibility due to the relationship of the parameters ϕ and τ in the mean and variance relationship, such that $Var(Y) = \phi\mu^{\tau}$. While the Tweedie distribution allows us to model the data exactly, the PV distribution provides an adequate approximation and is easily estimated through the use of the gamma probability density function. Moreover, this form allows us to analyze data without requiring the assumption that the data follow a known standard distribution. The motivation for this approximation arose from work in the analysis of exponential dispersion models with mean-variance relationships. More so, Jørgensen [12] proved that such distributions having the power variance form with $0 < \tau < 1$ cannot correspond to an exponential family distribution. This flexible formulation is useful in many practical applications.

The moments and properties of the gamma in its PV form are reproduced in Section 2. In Section 3, the PV form is compared to several known distributions in the exponential family. In Section 4, the model parameter estimation for the PV distribution using generalized least squares is described. A simulation study is conducted in Section 5. Two numerical examples are analyzed in Section 6, including applications in education and physical science. Some comments are given in Section 7.

2 Three-Parameter Gamma Distribution

In this section, we consider the gamma distribution in its power variance form and examine its properties under this formulation. Although this is the same parameterization as the Tweedie distribution [3], it differs in that the PV distribution is written in a closed form and thereby provides an approximation to distributions which can be described by the Tweedie distribution.

2.1 Density Function

Let Y denote a positive random variable with power variance of the form with the density function

$$f_Y(y;\mu,\phi,\tau) = \frac{y^{(\mu^{2-\tau}-\phi)/\phi}\, exp\,(-y\,\mu^{1-\tau}/\phi)}{\Gamma\left(\frac{\mu^{2-\tau}}{\phi}\right)\phi^{\frac{\mu^{2-\tau}}{\phi}}\,\mu^{(\tau-1)\mu^{2-\tau}/\phi}},\ y > 0 \tag{2.1}$$

within a parameter space $\Omega_\mu \times \Omega_\varphi \times \Omega_\tau \equiv \mathfrak{R}^+ \times \mathfrak{R}^+ \times \mathfrak{R}$. We postulate that the density function (2.1) represents random variables with the mean-variance relationship

$$Var(Y) = \phi\mu^{\tau},$$

where $\mu = E(Y)$, ϕ is a positive scale parameter, and τ is a real-valued parameter. The $f_Y(y;\mu,\phi,\tau)$ is a density function for the gamma distribution in a different form as it represents a re-parameterization of the two-parameter gamma $\Gamma(\alpha,\beta)$ with density function,

$$f_Y(y;\alpha,\beta) = \frac{y^{\alpha-1}\, exp\,(-y/\beta)}{\Gamma(\alpha)\,\beta^{\alpha}},\ y > 0,\ \ \alpha > 0,\ \ \beta > 0 \tag{2.2}$$

where the gamma distribution shape parameter $\alpha = \mu^{2-\tau}/\phi$ and the gamma distribution scale parameter $\beta = \phi\mu^{\tau-1}$ [13].

2.2 Moments and properties

The density function $f_Y(y;\mu,\phi,\tau)$ (2.1) has as its moment generating function

$$M_Y(t) = E(e^{tY}) = \int_0^\infty \frac{y^{\frac{\mu^{2-\tau}-\phi}{\phi}} exp\left(-y\left[\frac{\mu^{1-\tau}}{\phi}-t\right]\right)}{\Gamma\left(\frac{\mu^{2-\tau}}{\phi}\right)\phi^{\frac{\mu^{2-\tau}}{\phi}}\mu^{\frac{(\tau-1)\mu^{2-\tau}}{\phi}}} dy = (1-\phi\mu^{\tau-1}\,t)^{-(\mu^{2-\tau})/\phi}, \tag{2.3}$$

for $t < \mu^{1-\tau}/\phi$. From $M_Y(t)$, we obtain the cumulants of Y as

$$K_Y(t) = log[M_Y(t)] = -\left(\frac{\mu^{2-\tau}}{\phi}\right) log\,[1-\phi\,\mu^{\tau-1}t]. \tag{2.4}$$

In its general form, the r-th cumulant for the random variable Y has the form

$$\kappa_r = K_Y^{(r)}(0) = (r-1)!\,(\phi\,\mu^{\tau-1})^{r-1}\,\mu$$

for r = 1, 2, 3, … where $K_Y^{(r)}$ denotes the r-th derivative of K_Y with respect to the parameter τ. Thus, the random variable Y, has its mode at μ^{-1} $(\mu^2 - \phi\,\mu^\tau)$, provided that

$$Var(Y) = \phi\mu^\tau < \mu^2.$$

3 Comparison to the Exponential Family of Distributions

Let Y_i for $i = 1,2,\ldots n$ be a set of continuous n random variables. The family of distributions f_θ for $\theta \in \Theta$ belongs to the one parameter exponential family if the density of Y_i is of the form

$$f(y|\theta) = e^{\lambda(\theta)T(y)-\psi(\theta)}h(y) \tag{3.1}$$

for some real valued functions $\lambda(\theta)$, $T(y)$, $\psi(\theta)$, and $h(y) \geq 0$. The exponential family is a family of distributions on the finite dimensional Euclidean space, parameterized by a finite dimensional parameter vector. The family provides a framework that makes it useful and convenient in statistical analyses [14-17]. The exponential family contains as special cases most of the standard discrete and continuous distributions typically used for statistical modeling, such as the normal, Poisson, exponential, gamma, among others. The PV distribution is an extension of the gamma distribution. It is parameterized in the form (3.1) and is a member of the exponential family of distributions. We provide comparisons of the PV form to several common distributions in the exponential family as these distributions are widely used in modeling and have known mean-variance relationships.

3.1 Comparisons to the exponential, chi-square, and euler distributions

The PV distribution provides approximations to several known distributions. The most general family is the two-parameter gamma family, (2.2). The parameterization of the PV model is an alternative to the two-parameter representation of the gamma distribution [18]. They parameterized the $\Gamma(\alpha,\beta)$ distribution in terms of the mean and dispersion of the random variable as $\mu = \beta/\phi$ and $\phi = 1/\alpha$ respectively. The exponential and chi-square distributions are special cases of the two-parameter gamma family, and as such have an obvious representation in the PV distribution family. The Euler distribution is also a special case of the PV distribution. It has the form

$$f_Y(y;\,\mu,\phi) = \frac{y^{\frac{\mu}{\phi}-1} exp\,(-y/\phi)}{\Gamma\left(\frac{\mu}{\phi}\right)\,\phi^{\frac{\mu}{\phi}}}, \qquad y > 0, \qquad \mu > 0, \qquad \phi > 0,$$

with mean $E(Y) = \mu$ and variance $Var(Y) = \phi\mu^1$ which presents a special case of the PV distribution with $\tau = 1$. The parameter correspondences of the Euler, exponential, gamma, and chi-square distributions are summarized in Table 1.

Table 1. Parameter correspondence among PV_τ distribution and special cases

Distribution	Mean $E(Y)$	Variance $Var(Y)$	μ	ϕ	τ
PV	μ	$\phi\mu^\tau$	μ	ϕ	τ
Exponential	μ	μ^2	μ	1	2
Euler	μ	$\phi\mu$	μ	ϕ	1
Gamma	$\phi\mu$	$\phi\mu^2$	μ	ϕ	2
Chi-square	μ	2μ	μ	2	1

3.2 Mean-Variance relationship

Several exponential family distributions have the mean-variance relationship of the power form $Var(Y) = \phi\mu^\tau$ with the parameter τ restricted to integer values. We denote the distribution as the power variance distribution PV_τ with power parameter τ. For τ = 0, 1, 2, and 3, we have approximations of the Gaussian, Poisson, gamma, and inverse Gaussian distributions, respectively.

3.2.1 Gaussian distribution

Fig. 1 shows plots for the PV_0 (the PV distribution with power parameter $\tau = 0$) and the Gaussian densities with mean μ = 3 and different variances ($\phi\mu^\tau$). Although the Gaussian and PV distributions have the same means and variances in each of the three illustrations, it is clear that the distributions are quite different in their respective shapes as one nears the zero boundary of the PV distribution. The differences between the shapes of the two distributions become less pronounced as one moves further from zero or as the coefficient of variation, $\sigma(Y)/\mu(Y)$, decreases. In the three panels, Fig. 1, the means are equal to 3, while the variances are: (a) 3, (b) 1 and (c) 1/3. The red line represents the Gaussian curve, while the black line is the PV_0 curve. The coefficient of skewness, $\gamma_1 = E(Y-\mu)^3/Var(Y)^{3/2}$, for the PV_0 model approaches infinity as μ tends toward zero. As the mean μ tends towards infinity, the PV_0 distribution is similar to the Gaussian distribution as is shown in Fig. 1(c).

3.2.2 Poisson distribution

Fig. 2 provides a comparison of the PV_1 density and the Poisson mass function. In this illustration, the scale parameter ϕ was fixed to 1 so the two distributions have the same mean-variance relationships. The means are: (a) 2, (b) 5, and (c) 8. The red lines correspond to the Poisson mass function and the black line corresponds to the PV_1 density function. The plots show that the PV_1 distribution approximates the Poisson distribution as their relative shapes are similar, even when the mean approaches 0. The PV_1 distribution is slightly more skewed and platykurtic than the Poisson distribution. The differences become negligible as the means of the distributions become larger.

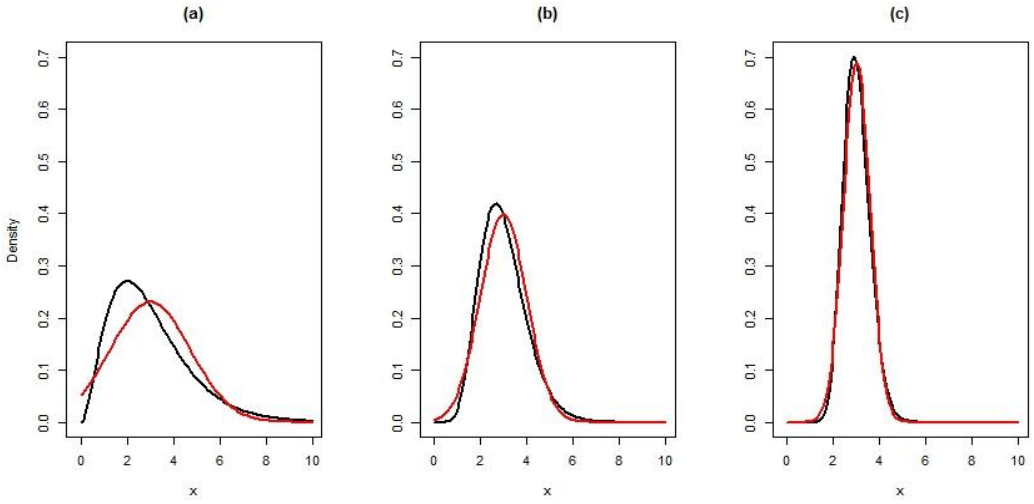

Fig. 1. Comparisons of the PV_0 distribution and the Gaussian distribution

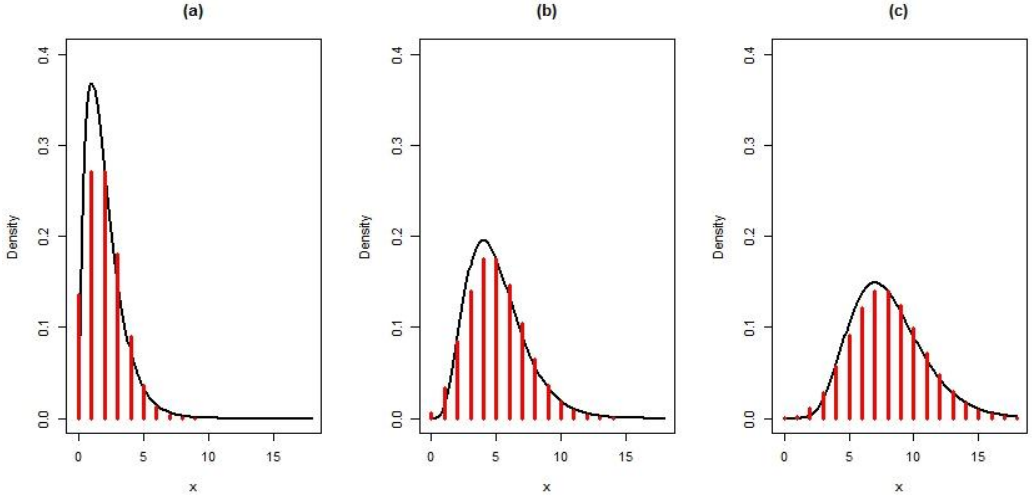

Fig. 2. Comparisons of the PV_1 distribution and the Poisson distribution

3.2.3 Log normal distribution

Fig. 3 shows plots of the PV_2 distribution and log normal $LN(\theta, \upsilon)$ distributions. We chose to use the log normal distribution because it is represented in the same power variance form with mean $E(Y) = e^{\theta+\upsilon/2} = \mu$ and variance $Var(Y) = (e^{\upsilon} - 1)\left(e^{\theta+\frac{\upsilon}{2}}\right)^2 = \phi\mu^2$. The means are equal to 4 in all panels, Fig. 3, while the scale parameters are: (a) ϕ = 1/4, (b) ϕ = 1/2 and (c) ϕ = 1. In Fig. 3, the red line is the log normal density and the black line is the PV_2 density. For these distributions, the scale parameter is the square of the coefficient of variation. The two distributions within each panel have the same mean and variance, although there are pronounced differences in the shapes of the distributions as the means approach zero. The two distributions are closer to each other when the coefficient of variation is small as in Fig. 3(a), compared to when the coefficient of variation is large as shown in Fig. 3(c).

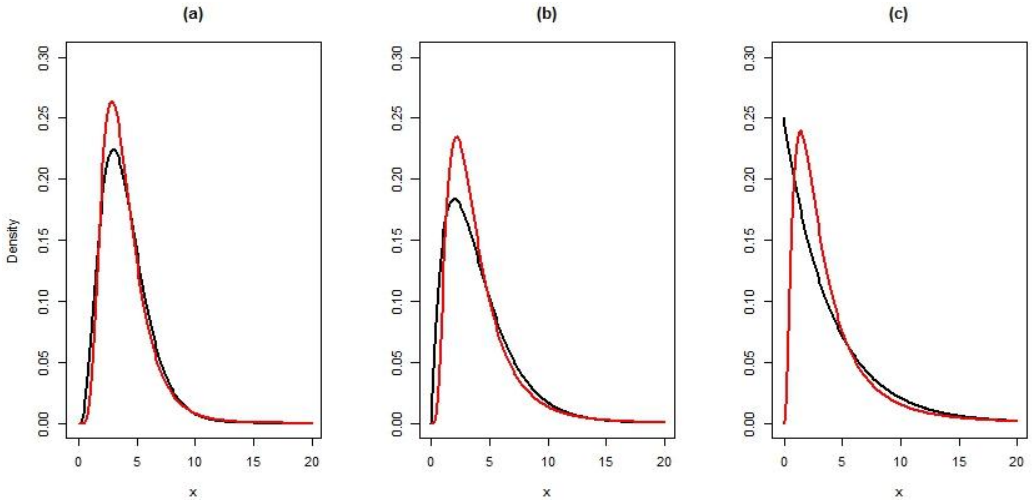

Fig. 3. Comparisons of the PV_2 distribution and the Log normal distribution

3.2.4 Moments of the PV distribution and comparative distributions

Table 2 provides higher-order moments and standardized moments for the PV distribution with various parameter values, as well as the Gaussian, Poisson, gamma, and inverse Gaussian distributions [19].

Table 2. Moments of the power variance and standard exponential family distributions

Distribution	$Var(Y)$	$E(Y-\mu)^3$	$E(Y-\mu)^4$	γ_1 [a]	γ_2 [b]
Normal	ϕ	0	$3\phi^2$	0	3
PV($\tau=0$)	ϕ	$2\phi^2\mu^{-1}$	$3\phi^2(1+2\phi\mu^{-2})$	$2\phi^{1/2}\mu^{-1}$	$3(1+2\phi\mu^{-2})$
Poisson	μ	μ	$3\mu^2\left(1+\frac{\mu^{-1}}{3}\right)$	$\mu^{-1/2}$	$3(1+\mu^{-1}/3)$
PV(μ,1,1)	μ	2μ	$3\mu^2(1+2\mu^{-1})$	$2\mu^{-1/2}$	$3(1+2\mu^{-1})$
Inverse Gaussian	$\phi\mu^3$	$3\phi^2\mu^5$	$3\phi^2\mu^6(1+5\phi\mu)$	$3\phi^{1/2}\mu^{1/2}$	$3(1+5\phi\mu)$
PV($\mu,\phi,3$)	$\phi\mu^3$	$2\phi^2\mu^5$	$3\phi^2\mu^6(1+2\phi\mu)$	$2\phi^{1/2}\mu^{1/2}$	$3(1+2\phi\mu)$
Euler[c]	$\phi\mu$	$2\phi^2\mu$	$3\phi^2\mu^2(1+2\phi\mu^{-1})$	$2\phi^{1/2}\mu^{-1/2}$	$3(1+2\phi\mu^{-1})$
Gamma[d]	$\phi\mu^2$	$2\phi^2\mu^3$	$3\phi^2\mu^4(1+2\phi)$	$2\phi^{1/2}$	$3(1+2\phi)$
PV(μ,ϕ,τ)	$\phi\mu^\tau$	$2\phi^2\mu^{2\tau-1}$	$3\phi^2\mu^{2\tau}(1+2\phi\mu^{\tau-2})$	$2\phi^{1/2}\mu^{\frac{\tau}{2}-1}$	$3(1+2\phi\mu^{\tau-2})$

[a] $\gamma_1=\frac{E(Y-\mu)^3}{(Var(Y))^{3/2}}$, *coefficient of skewness;* [b] $\gamma_2=\frac{E(Y-\mu)^4}{(Var(Y))^2}$, *standardized fourth moment;* [c] *Euler=* $PV(\mu,\phi,1)$; [d] *Gamma=* $PV(\mu,\phi,2)$

4 Generalized Least Squares Parameter Estimation

Due to the flexibility of the PV distributions, they are used to approximate other distributions. Thus, to estimate the parameters of the underlying distribution of the responses in the data, we present a regression

type estimator of these parameters from the mean variance relationship. Smyth [20] and Wilson and Koehler [21] among others have modeled dispersion based on multiple parameters as it pertains to different subpopulations instead with the use of a single constant factor. We obtain regression types estimators of the parameters ϕ and τ.

We consider a consistent estimator for the vector of parameters $\boldsymbol{\beta} = (\phi, \tau)'$ from a generalized least squares regression which allows the errors to have non-constant variance $Var(\epsilon) = \sigma^2\Omega$. The variance parameters are obtained through the linearized mean-variance relationship

$$log(Var(\boldsymbol{y})) = log\,\phi + \tau\, log\,\boldsymbol{\mu}.$$

For each subpopulation, we let the response vector $\boldsymbol{Y}$ has elements the log of the variance of the observed data with data matrix $\boldsymbol{X}$, consisting of a column of ones and a column of $log\,\mu$, is nonrandom with the vector of regression parameters $\boldsymbol{\beta}^* = (\beta_0, \beta_1)'$ where $\beta_0 = log\,\phi$ and $\beta_1 = \tau$. Group sizes are allowed to vary [22]. Thus, the estimates of our dispersion parameter and power parameter are obtained from an approximate generalized least squares estimator for $\boldsymbol{\beta}$,

$$\hat{\beta} = \left(X'\hat{\Omega}^{-1}X\right)^{-1}X\hat{\Omega}^{-1}Y$$

where the $\widehat{\boldsymbol{\Omega}}$ is a consistent estimator for the covariance matrix for $\boldsymbol{Y}$. The variance of $\widehat{\boldsymbol{\beta}}$ is $\sigma^2(\boldsymbol{X'\Omega^{-1}X})^{-1}$ [23,24].

The generalized least square estimates $\hat{\phi}$ and $\hat{\tau}$ are parameter estimates for the PV distribution, and regression coefficients for the data are obtained through minimizing the log likelihood of the PV distribution. Since the PV distribution is limited to positive values, values of $y = 0$ may be adjusted by a factor of 0.5 [25].

5 Simulation Study

The PV representation of the gamma distribution is flexible and can be used to approximate many distributions through the mean-variance relationship. We illustrate the applicability of the PV distribution through a simulation study. Data were simulated from a Poisson distribution Y with mean parameter μ, such that $log(\mu) = 1 + 2X$ where X is generated from an uniform (0, 1) distribution. We assume that the PV distribution is of form $\phi\mu^{\tau}$. The simulation was conducted for a total of 10,000 iterations, where the values of Y were predicted by:

a) Poisson distribution with a log link
b) PV distribution with a log link.

Parameter estimates were compared over three sample sizes (100, 250, 500) and four subgroup sizes (5, 10, 15, 20) to estimate the mean-variance relationship parameters ϕ and τ. The average generalized least squares estimates for the mean-variance parameters for each simulation condition are reported in Table 3.

In each simulation condition, the mean-variance parameters are estimated close to the true values of ϕ = 1 and τ = 1. Estimates of the power parameter τ improved as the number of subgroups in the data increased. The average regression parameter estimates and standard errors from each simulation condition are reported in Table 4. The model fit, measured through the mean square error (MSE), is also reported.

As expected, the Poisson distribution produced the most accurate parameter estimates. However, the PV distribution had comparable standard errors compared to the Poisson distribution. We note that the PV parameter estimates deviated further from the true values of β_0 = 1 and β_1 = 2 as the number of subgroups increased. Improved estimates were obtained with a larger number of observations in each subgroup (less subgroups) and larger sample sizes.

Table 3. Estimates of mean-variance parameters

Number of Subgroups	Sample Size	ϕ	τ
5	100	1.01	1.10
	250	0.95	1.10
	500	0.94	1.10
10	100	1.02	1.03
	250	0.98	1.03
	500	0.98	1.03
15	100	0.97	1.02
	250	0.97	1.02
	500	0.98	1.01
20	100	0.93	1.01
	250	0.97	1.00
	500	0.98	1.01

Table 4. Simulation results by number of subgroups and sample size

Number of Subgroups	Distribution	Sample Size	β_0 = 1.000		β_1 = 2.000		MSE
			Est	SE	Est	SE	
5	Poisson	100	0.997	0.092	2.002	0.130	8.470
		250	0.999	0.058	2.001	0.082	8.582
		500	0.999	0.041	2.000	0.058	8.641
	PV	100	0.951	0.089	2.063	0.129	8.514
		250	0.950	0.057	2.066	0.082	8.612
		500	0.950	0.040	2.067	0.058	8.666
10	Poisson	100	0.998	0.092	2.001	0.130	8.484
		250	1.000	0.058	2.000	0.082	8.596
		500	0.999	0.041	2.001	0.058	8.643
	PV	100	0.935	0.084	2.076	0.121	8.537
		250	0.941	0.055	2.072	0.078	8.628
		500	0.941	0.039	2.072	0.056	8.671
15	Poisson	100	0.997	0.092	2.003	0.130	8.492
		250	1.000	0.058	1.999	0.082	8.608
		500	0.999	0.041	2.001	0.058	8.639
	PV	100	0.924	0.082	2.086	0.116	8.558
		250	0.937	0.054	2.074	0.077	8.645
		500	0.938	0.039	2.074	0.055	8.668
20	Poisson	100	0.996	0.092	2.004	0.130	8.484
		250	0.999	0.058	2.000	0.082	8.589
		500	1.000	0.041	2.000	0.058	8.647
	PV	100	0.914	0.079	2.094	0.112	11.166
		250	0.931	0.053	2.079	0.076	8.630
		500	0.936	0.039	2.075	0.055	8.678

We compare the model fits of the PV distribution compared to the true distribution using the MSE. From the MSEs across the simulation conditions, we see that the PV distribution provides a comparable fit to the Poisson regression model in most cases. The model fit deviated the most between the two approaches in the case with a sample size of 100 and 20 subgroups. Under these conditions, there are only 5 observations per subgroup which can result in unreliable estimates for the mean-variance relationship parameters and impact the PV model fit. However, in all other simulation conditions, the PV model provided a good approximation.

6 Numerical Examples

We present two numerical examples demonstrating the use of the PV distribution in regression models for count and continuous data. We examine absenteeism among high school students and the compression strength of high-performance concrete. These examples demonstrate the use of the PV distribution as an approximation for different outcome types across various applications.

6.1 Absenteeism

Data were obtained from 316 9th-grade students enrolled in two high schools in the Los Angeles area in Fall 1995. Each student took the California Test of Basic Skills and received scores for the mathematics and language sections. We wish to explain the varying number of absences during the school year based on the student demographics and their academic performance as it pertains to ethnicity (a binary measure, Caucasian or Filipino versus other), the school, mathematics and language exam scores (percentile rank), and bilingual status.

The number of absences are evaluated using a generalized linear model with a Poisson distribution as the random component and a PV distribution model. We use the PV distribution as a flexible method to estimate the regression parameters and compare the results between the two approaches, Table 5. To estimate the mean-variance relationship parameters ϕ and τ based on the generalized least squares, we split the subpopulations into nine subgroups by ethnicity and school. We estimate the mean-variance parameters as $\hat{\phi}$ = 1.79 and $\hat{\tau}$ = 1.84, and obtain the PV model parameters by maximizing the log likelihood with parameters $\hat{\phi}$, $\hat{\tau}$, and $\hat{\mu}$.

Table 5. Parameter estimates, standard errors (SE), and *p* values for Poisson and PV models

Variables	Poisson			PV		
	Estimate	**SE**	***P***	**Estimate**	**SE**	***P***
Intercept	2.189	0.140	<.001	2.187	0.374	<.001
Ethnicity	0.679	0.079	<.001	0.758	0.186	<.001
School	-0.477	0.065	<.001	-0.451	0.175	.01
Mathematics Percentile Rank	-3.292×10^{-3}	1.319×10^{-3}	.01	-3.129×10^{-3}	3.602×10^{-3}	.39
Language Percentile Rank	3.298×10^{-5}	1.383×10^{-3}	.98	-1.200×10^{-3}	3.781×10^{-3}	.75
Bilingual	-0.212	0.049	<.001	-0.320	0.140	.02

The PV distribution allows additional flexibility in the model and provides a reasonable approximation to the Poisson model. The regression coefficient estimates are similar for both the Poisson and the PV models. We find that the model fit is similar for both models as the PV model has a MSE of 48.47 while the Poisson model has a MSE of 48.16.

6.2 Concrete compressive strength

High-performance concrete is a complex material which has increased compressive strength compared to conventional concrete. We are interested in predicting the strength of high-performance concrete, measured in megapascals, based on the content of 1,030 concrete samples [26]. The components (measured in kg/m^3) include cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate. The age of the concrete in days is also recorded. Using 14 subgroups based on the age of the concrete, the variance parameters were estimated as $\hat{\phi}$ = 17.77 and $\hat{\tau}$ = 0.35. The regression parameter estimates for both models are reported in Table 6.

Table 6. Parameter estimates, standard errors (SE), and *P* values for Gaussian and PV models

Variables	Gaussian			PV		
	Estimate	**SE**	***P***	**Estimate**	**SE**	***P***
Intercept	-23.331	26.586	.38	-23.337	18.499	.21
Cement	0.120	0.008	<.001	0.116	0.006	<.001
Blast Furnace Slag	0.104	0.010	<.001	0.093	0.007	<.001
Fly Ash	0.088	0.013	<.001	0.082	0.009	<.001
Water	-0.150	0.040	<.001	-0.128	0.028	<.001
Superplasticizer	0.292	0.093	.002	0.323	0.067	<.001
Coarse Aggregate	0.018	0.009	.05	0.014	0.007	.04
Fine Aggregate	0.020	0.011	.06	0.022	0.008	.004
Age	0.114	0.005	<.001	0.116	0.004	<.001

The MSEs for the Gaussian and PV models are 107.20 and 109.23, respectively, indicating that both approaches provide comparable model fits. The parameter estimates are fairly similar between the two models, although the PV approach produces smaller standard error estimates. The PV standard error estimates for the covariates are between 20 percent and 31 percent lower than the standard errors for the Gaussian model.

7 Conclusions

The PV form provides the flexibility to address the extra variation through the mean-variance relationship, as used in the Tweedie distribution, but easily estimated from a closed form of a gamma probability density function. We present the parameterization of the PV distribution and verify that it provides a good approximation of many other distributions described through the Tweedie distribution. We introduce a generalized least squares estimation method to estimate a two-parameter mean-variance relationship in ϕ and τ. Our simulation demonstrates the applicability of the PV distribution for estimating parameters in the data model from other distributions in the exponential family. In addition, the examples of modeling counts and continuous data pinpoint the flexibility that the PV distribution provides for a wide class of mean-variance relations.

This paper suggests that the PV distribution is useful to examine the performance of regression models for data that do not necessarily follow an exponential family but have a mean-variance relationship that may be related. While the PV distribution is used to simulate data that have a power form other than τ = 0, 1, 2, and 3, it is trivial to generate data that have a mean-variance relationship $Var(Y) = \phi\mu^{-1/2}$. Additionally, one can generate data with an alternative mean and variance relationships other than the power form. In particular, setting $\alpha = \mu^2/\phi V_\tau(\mu)$, and $\beta = \phi\mu V_\tau(\mu)$, yields a gamma model with a mean-variance relationship equal to $Var(Y) = \phi V_\tau(\mu)$ where $V_\tau(\mu)$ is the variance function indexed by the parameter τ [18]. It encompasses the negative binomial mean-variance relationship $Var(Y) = \mu(1 + \tau\mu)$ where the free scale parameter ϕ is set to unity. Similarly, one could generate data with mean $E(Y) = \pi$ and variance $Var(Y) = \phi\{\pi(1 - \pi)\}^\tau$. This manipulation of the mean-variance relationship through this PV model allows one to address overdispersed or underdispersed binomial.

The robustness of the quasi-likelihood function [27] and related methods, including the extended quasi-likelihood function [28] and the pseudo-normal likelihood [29] for non-exponential data, continues to be enhanced. Our discussion expands on this type of modeling.

Disclaimer

The work for this paper was conducted while the first author was at Arizona State University. The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Centers for Health Statistics, Centers for Disease Control and Prevention.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Cameron AC, Trivedi PK. Regression analysis of count data. New York: Cambridge Press; 1998.

[2] Dupont WD. Statistical modeling for biomedical researchers: A simple introduction to the analysis of complex data. New York: Cambridge Press; 2002.

[3] Tweedie MCK. An index which distinguishes between some important exponential families. In: Ghosh K, Roy J eds. Statistics: Applications and New Directions, Proceedings of the Indian Statistical Institute Golden Jubilee International Conference. Calcutta: Indian Statistical Institute; 1984.

[4] Anderson D, Feldblum S, Modlin C, Schirmacher D, Schirmacher E, Thandi N. A Practitioner's guide to generalized linear models. Colorado Springs: Casualty Actuarial Society Discussion Paper Program; 2004.

[5] Dunn PK, Smyth G. Series evaluation of Tweedie exponential dispersion model densities. Statistics and Computing. 2005;15(4):267-280.

[6] Dunn PK, Smyth G. Evaluation of Tweedie exponential dispersion model densities by Fourier inversion. Statistics and Computing. 2008;18(1):73-86.

[7] Bonat WH, Kokonendji CC. Flexible Tweedie regression models for continuous data. Journal of Statistical Computation and Simulation. 2017;87(11):2138-2152.

[8] Dunn PK. Tweedie: Tweedie exponential family models. R package version 2.1.7; 2013.

[9] Stacy EW. A generalization of the gamma distribution. The Annals of Mathematical Statistics. 1962;33:1187-1192.

[10] Stacy EW, Mihram GA. Parameter estimation for a generalized gamma distribution. Technometrics. 1965;7:349-358.

[11] Wilks DS. Comparison of three-parameter probability distributions for representing annual extreme and partial duration precipitation series. Water Resources Research. 1993;29:3543-3549.

[12] Jørgensen B. Exponential dispersion models (with discussion). Journal of the Royal Statistical Society Series B. 1987;49:127–162.

[13] Johnson NL, Kotz S. Distributions in statistics: Continuous distributions. New York: Wiley; 1970.

[14] Barndorff-Nielsen O. Information and exponential families in statistical theory. New York: Wiley; 1978.

[15] Brown LD. Fundamentals of statistical exponential families. Hayward, CA: IMS Lecture Notes and Monographs Series; 1986.

[16] Lehmann EL, Casella G. Theory of point estimation. New York: Springer; 1998.

[17] Bickel PJ, Doksum K. Mathematical statistics, basic ideas and selected topics, Vol I. Saddle River: Prentice Hall; 2006.

[18] McCullagh P, Nelder JA. Generalized linear modelS, 2nd ed. London: Chapman and Hall; 1989.

[19] Casella G, Berger RL. Statistical inference, 2nd Edition. Pacific Grove: Duxbury; 2002.

[20] Smyth GK. Generalized linear models with varying dispersion. Journal of the Royal Statistical Society Series B. 1989;51(1):47-60.

[21] Wilson JR, Koehler KJ. Hierarchical models for cross-classified overdispersed multinomial data. Journal of Business and Economic Statistics. 1991;9(1):103-110.

[22] Lemeshow S, Hosmer DW. A review of goodness of fit statistics for use in the development of logistic regression models. American Journal of Epidemiology. 1982;115(1):92-106.

[23] Johnston J. Econometric methods. New York: McGraw-Hill; 1972.

[24] Amemiya T. Advanced econometrics. Cambridge: Harvard University Press; 1985.

[25] Bishop Y, Fienberg SE, Holland P. Discrete multivariate analysis: Theory and practice. Cambridge: MIT Press; 1975.

[26] Yeh I. Modeling of strength of high performance concrete using artificial neural networks. Cement and Concrete Research. 1998;28(12):1797-1808.

[27] Wedderburn RWM. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika 1974;61:439–447.

[28] Nelder JA, Pregibon D. An extended quasi-likelihood function. Biometrika. 1987;74:221–232.

[29] Carroll RJ, Ruppert D. Robust estimation in heteroskedastic linear models. Annals of Statistics. 1982;10:429–441.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sciencedomain.org/review-history/28042