Essays on the Identification and Modeling of Variance

by

Katherine E. Irimata

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved March 2018 by the
Graduate Supervisory Committee:

Jeffrey R. Wilson, Chair
Ioannis Kamarianakis
Ming-Hung Kao
Mark Reiser
John Stufken

ARIZONA STATE UNIVERSITY

May 2018

ABSTRACT

In the presence of correlation, generalized linear models cannot be employed to obtain regression parameter estimates. To appropriately address the extravariation due to correlation, methods to estimate and model the additional variation are investigated. A general form of the mean-variance relationship is proposed which incorporates the canonical parameter. The two variance parameters are estimated using generalized method of moments, negating the need for a distributional assumption. The mean-variance relation estimates are applied to clustered data and implemented in an adjusted generalized quasi-likelihood approach through an adjustment to the covariance matrix. In the presence of significant correlation in hierarchical structured data, the adjusted generalized quasi-likelihood model shows improved performance for random effect estimates. In addition, submodels to address deviation in skewness and kurtosis are provided to jointly model the mean, variance, skewness, and kurtosis. The additional models identify covariates influencing the third and fourth moments. A cutoff to trim the data is provided which improves parameter estimation and model fit. For each topic, findings are demonstrated through comprehensive simulation studies and numerical examples. Examples evaluated include data on children's morbidity in the Philippines, adolescent health from the National Longitudinal Study of Adolescent to Adult Health, as well as proteomic assays for breast cancer screening.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

While many statistical tests and models focus on understanding and characterizing the mean, the variance has an important role. Many distributions make assumptions about the variance, which must be met when analyzing or modeling data. Evaluating deviations in the assumed variance can help researchers to better understand their data and verify the appropriateness of the selected model. Moreover, modeling the variance such as in weighted least squares or joint modeling improves model fit and accounts for excess variability. Improvements to modeling techniques have reduced the variance or the uncertainty in understanding the driver for the model. Adjustments to estimation and modeling methods to account for the variance can be implemented to maximize the information available.

1.1 Exponential Family of Distributions

A random variable Y has a probability distribution which is a member of the exponential family if the density function can be written in the form

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{\left(y\theta - b(\theta)\right)}{a(\phi)} + c(y,\phi)\right\}$$

for canonical parameter θ, and functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$. Many common distributions, such as the normal, Poisson, gamma, and binomial distributions can be characterized in this form. Distributions in the exponential family have the following properties: the mean is $E(Y) = \mu = b'(\theta)$ and the variance is $var(Y) = b''(\theta)a(\phi)$. The variance, composed of a function of $b(\cdot)$, is related to the mean and can also be expressed in terms of the mean with the function $V(\mu)$ (McCullagh and Nelder 1989).

1.2 Moments of Distributions

A probability distribution has the moment generating function

$$M_Y(t) = E(e^{tY})$$

where t is in the real numbers. Moment generating functions are used to obtain properties of random variables and can be used to characterize distributions. For a continuous random variable y, the first moment or mean of a distribution can be found as

$$M_y'(0) = \frac{dM_y(t)}{dt}\bigg|_0 = \left[\int_{-\infty}^{\infty} ye^{ty}f(y)dy\right]_0 = \int_{-\infty}^{\infty} yf(y)dy = E(y).$$

Higher order moments are determined by taking the corresponding derivative of the moment generating function and evaluating the function at $t = 0$ (Rencher and Schaalje 2008).

The mean and the variance, the first and second central moments, are often used to describe a distribution. The mean identifies the center of the distribution, while the variance describes the spread. The measures skewness and kurtosis are also used to describe the shape of the distribution. These measures are related to the third and fourth central moments where the skewness is defined as

$$\gamma_1 = \frac{E(Y-\mu)^3}{(\sigma^2)^{3/2}}$$

and the excess kurtosis is defined as

$$\gamma_2 = \frac{E(Y-\mu)^4}{(\sigma^2)^2} - 3,$$

for μ, the mean of Y, and the variance σ^2. The skewness and kurtosis both describe the tails of the distribution. The skewness measures symmetry in the tails of the distribution, while the kurtosis indicates the weight of the tails.

1.3 Generalized Linear Models

Generalized linear models are an extension of linear models, and allow for the response distribution to be a member of the exponential family (not necessarily normal) and do not require the relationship between the response and the predictors to be linear. Generalized linear models are specified by three components: the random component, systematic component, and the link function. The random component identifies the probability distribution of the response variable Y. The systematic component describes the linear predictor η, which is a function of the covariates so that $\eta = \sum_{j=1}^{p} x_j \beta_j$. The link function specifies the relationship between the random and systematic components through the function $g(\cdot)$ which is a monotonic twice differentiable function. Thus for μ, the expected value of Y, $g(\mu) = \eta$ (McCullagh and Nelder 1989).

1.4 Summary of Chapters

In this dissertation, three distinct papers are presented. The first paper explores the estimation of the mean-variance relationship, particularly in the presence of overdispersion and underdispersion. An alternative parameterization using the canonical parameter is proposed, and the mean-variance parameters are estimated using generalized method of moments. This approach expands the variance form to account for overdispersion, as suggested by the data.

The second paper utilizes the estimation of the mean-variance relation and extends it for use in two-level clustered data. An adjusted generalized quasi-likelihood modeling approach is presented that implements the estimated variance in the covariance matrix. This adjustment allows generalized quasi-likelihood models to account for the

true mean-variance relationship prescribed by the data and offers a method for modeling overdispersed data.

The third paper further evaluates distributional assumptions through higher order moment conditions, including the skewness and kurtosis. Joint modeling of the mean and dispersion (Smyth 1989) is expanded to incorporate deviation in skewness and deviation in skewness submodels. These models identify covariates related to deviations in the tails of the distribution. A cutoff is also investigated to remove outliers due to skewness and kurtosis and improve model fit for the mean and dispersion submodels.

CHAPTER 2

ADJUSTED CANONICAL PARAMETER FOR ESTIMATING THE MEAN-VARIANCE RELATIONSHIP

Abstract

Implicit to parametric modeling is an assumption of the underlying properties of the distribution, which typically includes the specification of the relationships between the mean and variance. While many distributions in the exponential family have a theoretical mean-variance relationship, it is often the case that the data under investigation are correlated thus varying the relation. While others have adjusted the likelihood to estimate the true mean-variance relationship, we present a generalized method of moments approach based on an adjustment to the canonical parameter as warranted by the data. This method is void of distributional assumptions and is computationally tractable. We provide test statistics and confidence intervals for identifying the mean-variance parameters relation. The adjustment through the canonical parameter provides a general approach for all models with unknown underlying distributions but with memberships in the quasi-exponential family. The properties and performance of our method are evaluated through a simulation study. Two numerical examples were analyzed, one with a count outcome and the other with a binary outcome.

2.1 Introduction

The construction of test statistics and confidence intervals rely on the variance of the responses. As a common statistical measure, the variance is often relied on to

understand the differences in the responses or a function of the responses for a set of data. We often make assumptions about the variance based on the assumed underlying distribution of the responses, as it is well known that the variance is related to the mean for most distributions in the exponential family. However, it is often the case that while the responses may be on a certain scale and often resemble a certain distribution, the mean-variance relationship may not be as expected due to extraneous variation. This extraneous variation may be due in part to clustering or the structural design of the data collected. As such, the true but unknown distribution has parameters that are shifted.

A shift in parameter values, indicating that the true variance of the responses deviates from the expected variance, is commonly exhibited in count and binary data. Inflated variance, or so-called overdispersion, is often observed, particularly in longitudinal or clustered data analyses, as overdispersion is inherent in data with a hierarchical structure. McCullagh and Nelder (1989) suggested that overdispersion may be the norm. While underdispersion describes the case when the variance is a fraction of the expected mean, this is less often reported in practice.

Overdispersion has two primary effects. One is that the summary statistics, including the test statistics, will have a larger variance than expected (Morel and Neerchal 2012). The second effect is a possible loss of efficiency in using statistics appropriate for the single-parameter family and ignoring the variance (Cox 1983). Therefore, it is important to identify and account for any overdispersion; otherwise, one is likely to declare covariates as significant when in fact they are not. Studies have also shown that ignoring overdispersion and thereby misspecifying the model can bias the covariate effects and greatly impact the variance of the coefficients (Wilson and Koehler

1991; Milanzi, Alonso, and Molenberghs 2012). While underdispersion is less common, it also impacts the accuracy of the analysis and can result in inaccurate conclusions.

Methods to identify overdispersion and provide corrections to improve estimates of the variance have been presented (Cox and Reid 1987; McCullagh and Tibshirani 1990). In addition, several methods to address overdispersion due to correlation have been widely investigated. Some methods account for correlation through the random component, through analyses such as generalized estimating equations (Liang and Zeger 1986) or by assuming a probability distribution on the response. Other studies have considered tests for specific distributions, such as tests for overdispersion for proportions (Pack 1986). Score test statistics for Poisson and binomial models with overdispersion have also been presented (Dean 1992), with extensions to results for more general distributions. Xiang et al. (2007) provided a score test for overdispersion in a zero-inflated Poisson mixed regression model. Yang, Hardin, and Addy (2009) simplified the score statistic to test overdispersion in the zero–inflated generalized Poisson mixed model which was selected based on the approximate mean-variance relationship in the data.

Distributions in the exponential family exhibit a certain mean-variance relationship. This relation is altered if there is overdispersion or underdispersion. As such, overdispersion or underdispersion is identified by estimating the parameters in the mean-variance relationship and then measuring deviations from the theoretical values. Mean-variance models are not uncommon, in fact, Kukush et al. (2009) considered a pair of mean and variance functions with a common parameter vector θ estimated using an extended quasi-score function. Tsou (2011) considered two parameters (λ, ψ) in a parametric robust method of determining the mean-variance relationship through

estimation of the power λ with an adjusted robust log likelihood method for fixed values of ψ. Tsou demonstrated that, for distributions with a power mean-variance relationship such as the Poisson, Weibull, and Inverse Gaussian, the adjusted log likelihood method produced accurate parameter estimates despite using a normal working model. While these methods produce good results, the performance does not extend to nonlinear relationships, such as with binary data, which does not have a power relationship between the mean and variance.

This paper presents the evaluation of a general mean-variance relation in observational data. It consists of a simple method to identify and estimate overdispersion and underdispersion based on a two-parameter representation of the mean-variance relationship. We present a generalized method of moments (GMM) approach based on an adjusted canonical parameterization, which generalizes this method to all distributions that are members of the quasi-exponential family. Our approach negates the need for distributional assumptions as required with a maximum likelihood estimation approach. In addition, we provide test statistics and confidence intervals for the adjustment parameters. The properties and performance of our estimators are validated through a simulation study.

In Section 2.2, we review a likelihood estimation approach for the two-parameter mean-variance relation. In Section 2.3, we introduce the canonical parameterization of the mean-variance relationship, which expands the use of this variance form to any distribution in the exponential family. We present a generalized method of moments approach to estimate the mean-variance parameters as well as a test to identify overdispersion or underdispersion. In Section 2.4, a simulation study is conducted to

examine the properties of the GMM estimators. In Section 2.5, our GMM approach of estimating the mean-variance relationship is applied to a data set obtained from Weisberg (1985), the count of the number of snow geese in a flock, and to a children's morbidity study in the Philippines (Bhargava 1994).

2.2 Likelihood Approach to Estimating the Mean-Variance Relationship

Consider n observations $y_1, \dots, y_n$ of the random variable $\boldsymbol{Y}$, each with mean μ_i and variance

$$var(y_i) = \psi\mu_i^{\lambda}.$$

Tsou (2011) provided a parametric robust method of determining this mean-variance relationship by estimating the power λ with an adjusted robust log likelihood method for fixed values of ψ. He proposed adding a correction to the normal model to obtain asymptotically valid inferences for λ, ψ, and the regression parameters. For the normal distribution, the log profile likelihood for λ is

$$l\big(\lambda, \pi(\lambda)\big) = -\frac{1}{2} log\ 2\pi - \frac{1}{2} log\ \psi\mu^{\lambda} - \frac{(y-\mu)^2}{2\psi\mu^{\lambda}},$$

to which Tsou implements a robust adjustment $\frac{A}{B}$ (Royall and Tsou 2003). The adjustment factors are

$$A = \boldsymbol{I}_{\lambda\lambda} - \boldsymbol{I}_{\lambda\pi}\boldsymbol{I}_{\pi\pi}^{-1}\boldsymbol{I}_{\pi\lambda}$$

and

$$B = \boldsymbol{V}_{\lambda\lambda} - 2\boldsymbol{I}_{\lambda\pi}\boldsymbol{I}_{\pi\pi}^{-1}\boldsymbol{V}_{\pi\lambda} + \boldsymbol{I}_{\lambda\pi}\boldsymbol{I}_{\pi\pi}^{-1}\boldsymbol{V}_{\pi\pi}\boldsymbol{I}_{\pi\pi}^{-1}\boldsymbol{I}_{\pi\lambda},$$

where λ is the power, $\boldsymbol{\pi} = \big(\beta_{I,0}, \dots, \beta_{I,p-1}, \psi\big)$ denotes the dispersion parameter and the regression coefficients, and $\boldsymbol{V}_{\lambda\lambda}$, $\boldsymbol{V}_{\lambda\pi}$, and $\boldsymbol{V}_{\pi\pi}$ are defined as follows:

$$\boldsymbol{V}_{\lambda\lambda} = \lim_{n\to\infty} E_h\left[\frac{l_{\lambda}(\lambda_0,\pi_0)^2}{n}\right],$$

$$\boldsymbol{V}_{\lambda\pi} = \lim_{n\to\infty} E_h\left[\frac{l_\lambda(\lambda_0,\pi_0)l_\pi(\lambda_0,\pi_0)}{n}\right],$$

$$\boldsymbol{V}_{\pi\pi} = \lim_{n\to\infty} E_h\left[\frac{l_\pi(\lambda_0,\pi_0)^2}{n}\right].$$

The terms l_λ and l_π are the first derivatives of $l\big(\lambda,\pi(\lambda)\big)$ in terms of λ and π, respectively. The limiting values of the derivatives of square matrix of dimension $p+1$,

$$\boldsymbol{I}_{\lambda\lambda} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{(log\,\mu_i)^2}{2},$$

and

$$\boldsymbol{I}_{\lambda\pi} = \lim_{n\to\infty} \frac{1}{n}\left[\sum_{i=1}^{n} \frac{\lambda x_{i,0}\mu_i'\,log\,\mu_i}{2\mu_i}, \dots, \sum_{i=1}^{n} \frac{\lambda x_{i,p-1}\mu_i'\,log\,\mu_i}{2\mu_i}, \sum_{i=1}^{n} \frac{log\,\mu_i}{2\psi}\right],$$

where the jth row for $j = 1, \dots, p$ is

$$\lim_{n\to\infty} \frac{1}{n}\left(\sum_{i=1}^{n} x_{i,j-1}x_{i,0}\left[\frac{{\mu_i'}^2}{\psi\mu_i^\lambda} + \frac{\lambda^2{\mu_i'}^2}{2\mu_i^2}\right], \dots, \sum_{i=1}^{n} x_{i,j-1}x_{i,p-1}\left[\frac{{\mu_i'}^2}{\psi\mu_i^\lambda} + \frac{\lambda^2{\mu_i'}^2}{2\mu_i^2}\right], \sum_{i=1}^{n}\left[\frac{\lambda x_{i,j-1}\mu_{i'}}{2\psi\mu_i}\right]\right),$$

and the $(p+1)^{\text{th}}$ row is

$$\lim_{n\to\infty} \frac{1}{n}\left(\sum_{i=1}^{n}\left[\frac{\lambda x_{i,0}\mu_i'}{2\psi\mu_i}\right], \dots, \sum_{i=1}^{n}\left[\frac{\lambda x_{i,p-1}\mu_i'}{2\psi\mu_i}\right], \frac{n}{2\psi^2}\right).$$

Tsou (2011) demonstrated that, for a fixed multiplicative factor ψ, the adjusted log likelihood method has good properties and the estimates for the power relating the mean and variance were produced despite using a normal working model. Tsou's simulation study showed that the power λ was estimated nearly exactly for the Poisson, Weibull, and Inverse Gaussian distributions with large sample sizes. In addition, the likelihood technique corrected the empirical type I error probability. While the robust method produced accurate estimates of the mean-variance relationship parameters, the adjustment factor requires lengthy derivations for the derivatives and depends on the skewness and the kurtosis of the true underlying distribution. In practice, this method is

computationally intensive to obtain the estimate for λ, and information about the underlying distribution is often unknown.

We present an alternative approach based on generalized method of moments. This estimation technique is based on the first and second moment conditions, assuming that they exist (Tan, et al. 2010). As such, this method does not require the distributional assumptions of likelihood methods and is computationally very tractable. Our GMM method produces consistent estimates, similar to semiparametric techniques.

2.3 Generalized Method of Moments Approach to Estimating the Mean-Variance Relationship

2.3.1 General Framework

Consider n observations y_i, $i = 1, 2, \dots n$; as realizations of a set of independent random variables Y_i with mean μ_i related to k covariates $x_1, \dots, x_k$ through a link function $g(\cdot)$ so that $E(Y_i) = \mu_i$ and $g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k}$ for $i = 1, \dots, n$. The estimates of the regression parameters $\beta_0, \dots, \beta_k$ can be obtained using estimating equations obtained from the likelihood. Let the joint probability function of Y_i be

$$f(y; \theta, \phi) = exp\left[\frac{\left(y\theta - b(\theta)\right)}{a(\phi)} + c(y, \phi)\right]$$

with known functions a, b, and c (McCullagh and Nelder 1989). When the so-called dispersion parameter ϕ is known, then $f(y; \theta, \phi)$ is a linear exponential family model with canonical parameter θ. If θ is unknown, then we have a two-dimensional exponential family with log likelihood,

$$l(\theta, \phi | y) = \frac{\left(y\theta - b(\theta)\right)}{a(\phi)} + c(y, \phi).$$

Using the expectation of differentiation of the likelihood

$$E\left(\frac{\partial l(\theta,\phi|y)}{\partial\theta}\right)=0$$

and the property

$$E\left(\frac{\partial^2 l(\theta,\phi|y)}{\partial\theta^2}\right)+E\left(\frac{l(\theta,\phi|y)}{\partial\theta}\right)^2=0$$

yields the expected value $E(Y)=b'(\theta)$ and the variance $var(Y)=b''(\theta)a(\phi)$ where $b'(\theta)$ denotes the first derivative and $b''(\cdot)$ denotes the second derivative. Thus, both b' and b'' are functions of the canonical parameter θ. The mean and the variance are related through the first derivative and second derivative of the function $b(\theta)$. The variance of the observations is a product of a function of the canonical parameter θ and a function of the dispersion parameter ϕ.

The Poisson distribution and the binomial distribution are members of the exponential family and are commonly used to analyze count data and binary data, respectively. The Poisson distribution has probability mass function

$$f(y;\alpha)=exp(y\,log\,\alpha-\alpha-log\,y!)$$

with $a(\phi)=1$, $b(\theta)=\alpha=exp(\theta)$, $c(y,\phi)=-log\,y!$, and canonical parameter $\theta=log(\alpha)$. Thus, the expected mean and variance under the Poisson distribution are equal to α. The binomial distribution has probability distribution function

$$f(y;m,p)=\binom{m}{y}p^y(1-p)^{m-y},$$

with $a(\phi)=1$, $b(\theta)=mlog(1+exp(\theta))$, $c(y,\phi)=log\binom{m}{y}$ and the canonical parameter $\theta=log\left(\frac{p}{1-p}\right)$. Under the binomial distribution, the expected mean is mp and

the expected variance is $mp(1-p)$ (McCullagh and Nelder 1989; Dobson and Barnett 2008).

2.3.2 Canonical Parameterization

Consider the canonical parameter θ through its derivative of the inverse link function h, where $h = g^{-1}$, and ψ and λ are parameters in the variance of Y. Then, a general mean-variance relationship with flexibility through the parameters of the dispersion ψ and power λ, is

$$var(Y) = \psi\left[h'\left(g(\mu)\right)\right]^{\lambda} = \psi[h'(\theta)]^{\lambda} = b''(\theta)a(\phi)$$

where $h'(\theta)$ is the first derivative of the inverse of the canonical link. For a majority of members of the exponential family of distributions, including the Poisson and binomial distributions, this power relationship is convenient and flexible to describe the mean-variance relationship.

For example, in the case where Y follows the Poisson distribution with natural parameter α, the canonical parameter is $\theta = log(\alpha)$ and so the corresponding mean-variance parameter relation is

$$var(Y) = \psi[h'(\theta)]^{\lambda} = \psi\alpha^{\lambda}$$

which reflects an adjusted true mean-variance parameter relation. For the binomial distribution with natural parameters m and p, the canonical parameter is $\theta = logit(p)$ which corresponds to the canonical mean-variance parameter relation

$$var(Y) = \psi[h'(\theta)]^{\lambda} = \psi\left[\frac{exp(\theta)}{(1+exp(\theta))^2}\right]^{\lambda} = \psi[p(1-p)]^{\lambda},$$

as $p = e^{\theta}\left(1+e^{\theta}\right)^{-1}$. Thus, the power parameter λ allows for adjustment or deviation from the specified distributional properties. While the binomial does not have a natural

power relationship between the mean and variance, the relationship based on the canonical parameter can be estimated. We consider such deviation while considering distributions from the quasi-exponential family. The quasi-exponential families represents the Poisson and the binomial distributions but do not possess a full identification of the true underlying distribution. They are fully defined through the canonical parameterization identified based on the scale of the responses in the data. Such situations are robust to misspecification as shown in Section 2.4. In this paper, we introduce GMM estimators for the parameters ψ and λ in the canonical mean-variance relation.

2.3.3 GMM Estimation of ψ and λ

The parameters, ψ and λ, are key parameters in the variance function and so it is essential that we obtain reliable and efficient estimates. Our GMM estimators identify deviations from the theoretical values in the true mean-variance relation. We rely on the assumptions that the distribution is a member of the quasi-exponential family and that the first and second moments exist (Hansen 1982). We do not require complete distributional assumptions, as is the case with likelihood estimation. However, the GMM estimators of ψ and λ are consistent and are asymptotically normal (Jiang 2003).

Let $\widehat{\boldsymbol{\gamma}}_{GMM}$ be an estimator for a vector of parameters $\boldsymbol{\gamma} = (\psi, \lambda)'$ that minimizes the quadratic objective function $f_n(\boldsymbol{\gamma})' W_n f_n(\boldsymbol{\gamma})$ (Zsohar 2012; Lalonde, Wilson, and Yin 2014), where $f_n(\boldsymbol{\gamma})$ is a vector of the sample moment conditions, and W_n is a symmetric, positive definite weight matrix of dimension n. Then,

$$\widehat{\boldsymbol{\gamma}}_{GMM} = argmin_{\boldsymbol{\beta}}\{f_n(\boldsymbol{\gamma})' W_n f_n(\boldsymbol{\gamma})\} \tag{2.1}$$

is a the generalized method of moments estimator for $\boldsymbol{\gamma}$ which minimizes the objective function. Thus, we obtain the GMM estimators of the parameters ψ and λ, by presenting the population moment conditions

$$E\left(h'(\theta_i)\left(var(y_i) - \psi[h'(\theta_i)]^{\lambda}\right)\right) = 0 \tag{2.2a}$$

$$E\left(h'(\theta_i)^2\left(var(y_i) - \psi[h'(\theta_i)]^{\lambda}\right)\right) = 0 \tag{2.2b}$$

where $h'(\theta_i)$ is the first derivative of the inverse link function. These conditions are similar to moment conditions for estimation of parameters in a nonlinear model, where $\psi[h'(\theta_i)]^{\lambda}$ as an unbiased estimate of the variance and $var(y_i)$ as the empirical estimate of the variance based on the data, $(y_i - \mu_i)^2$. Equating the moment condition and an empirical estimate of $f_n(\boldsymbol{\gamma})$ results in

$$\frac{1}{n}\sum_{i=1}^{n} f(y_i, \psi, \lambda) = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} h'(\theta_i)\left(var(y_i) - \psi[h'(\theta_i)]^{\lambda}\right) \\ \frac{1}{n}\sum_{i=1}^{n} h'(\theta_i)^2\left(var(y_i) - \psi[h'(\theta_i)]^{\lambda}\right) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

A two-step GMM is utilized, with an identity weight matrix for the first iteration. In the second step, the weights are selected as an estimate of the optimal weight matrix for GMM as

$$\widehat{W}_n = \left[\frac{1}{n}\sum_{i=1}^{n} f\left(y_i, \hat{\psi}, \hat{\lambda}\right) f\left(y_i, \hat{\psi}, \hat{\lambda}\right)'\right]^{-1}$$

where $\hat{\psi}, \hat{\lambda}$ is an estimate of the mean-variance relationship from the first step (Imbens and Spady 2002). Thus, $\widehat{\boldsymbol{\gamma}}_{GMM}$ is an estimate of the mean-variance relation parameters that minimizes the quadratic objective function $f_n(\boldsymbol{\gamma})' W_n f_n(\boldsymbol{\gamma})$.

The generalized method of moments approach is flexible and estimates both parameters (ψ, λ) simultaneously rather than requiring one value to be held fixed while

the other is estimated. However, one can fix one parameter at a time and estimate the other parameter using one moment condition. Similarly, one can extend GMM and use the additional moment condition,

$$f_3(y_i, \psi, \lambda) = h'(\theta_i)^3\left(var(y_i) - \psi[h'(\theta_i)]^\lambda\right) \tag{2.3}$$

to estimate parameters. The additional moment condition improves the asymptotic efficiency, although there is the possibility of small sample bias (Donald, Imbens, and Newey 2009). This GMM procedure has many advantages over the likelihood approach. It does not rely on complete distributional assumptions, and is obtainable even when likelihood methods are computationally burdensome (Zsohar 2012). In addition, the resulting estimates for the variance parameters ψ and λ are reliable and consistent.

2.3.4 Inference for the Mean-Variance Relation

We obtain the asymptotic properties of our estimators, develop a test of overdispersion, and obtain the confidence intervals for the parameters through a GMM approach (Hansen 1982). Assume the data come from a quasi-exponential family. Since the sample moments are asymptotically normally distributed, we have

$$\sqrt{n}\left(f_n(\widehat{\boldsymbol{\gamma}})\right) \xrightarrow{d} N(0, \boldsymbol{\Delta}),$$

where

$$\boldsymbol{\Delta} = E[f(y, \boldsymbol{\gamma}^*)f(y, \boldsymbol{\gamma}^*)']$$

for a value of the parameter $\boldsymbol{\gamma}^*$. As such, the GMM estimator $\widehat{\boldsymbol{\gamma}}_{GMM}$ has the asymptotic covariance,

$$var(\widehat{\boldsymbol{\gamma}}_{GMM}) = \boldsymbol{V}_{GMM} = \frac{1}{n}[\boldsymbol{\Gamma}'\boldsymbol{W}\boldsymbol{\Gamma}]^{-1}\boldsymbol{\Gamma}'\boldsymbol{W}\boldsymbol{\Delta}\boldsymbol{W}\boldsymbol{\Gamma}[\boldsymbol{\Gamma}'\boldsymbol{W}\boldsymbol{\Gamma}]^{-1}$$

for

$$\boldsymbol{\Gamma} = E\left[\frac{\partial f(y,\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\right] = E\left[\frac{\partial f(y,\lambda,\psi)}{\partial \psi}, \frac{\partial f(y,\lambda,\psi)}{\partial \lambda}\right]',$$

where $\boldsymbol{\Gamma}$ is the expected value of the Jacobian of population moment conditions and $\boldsymbol{W}$ is a specified weight matrix. In the optimal case, the weight matrix is selected as $\boldsymbol{W} = (\boldsymbol{\Delta})^{-1}$, so that

$$\boldsymbol{V}_{GMM} = \frac{1}{n}[\boldsymbol{\Gamma}'\boldsymbol{W}\boldsymbol{\Gamma}]^{-1}$$

resulting in asymptotically efficient GMM estimators, $\hat{\psi}_{GMM}$ and $\hat{\lambda}_{GMM}$ (Zsohar 2012). In practice, the covariance matrix is calculated using the estimate

$$\widehat{\boldsymbol{\Gamma}} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \frac{\partial f(y,\widehat{\boldsymbol{\gamma}})}{\partial \widehat{\boldsymbol{\gamma}}}$$

for $\widehat{\boldsymbol{\gamma}} = (\hat{\psi}, \hat{\lambda})$. Significant overdispersion is identified through testing the hypotheses $H_0: \psi = 1$, $H_a: \psi > 1$ and $H_0: \lambda = 1$, $H_a: \lambda > 1$. The Z test statistics

$$Z_{\psi} = \frac{\hat{\psi}-1}{\sqrt{var(\hat{\psi})}}$$

and

$$Z_{\lambda} = \frac{\hat{\lambda}-1}{\sqrt{var(\hat{\lambda})}},$$

and follow the standard normal distribution under the null hypothesis. Thus, a measure of the overdispersion can be obtained through the 100(1-α)% confidence intervals for ψ and λ,

$$\left(\hat{\psi}_{GMM} - z_{1-\frac{\alpha}{2}}\sqrt{V_{GMM,\psi}}, \hat{\psi}_{GMM} + z_{1-\frac{\alpha}{2}}\sqrt{V_{GMM,\psi}}\right)$$

$$\left(\hat{\lambda}_{GMM} - z_{1-\frac{\alpha}{2}}\sqrt{V_{GMM,\lambda}}, \hat{\lambda}_{GMM} + z_{1-\frac{\alpha}{2}}\sqrt{V_{GMM,\lambda}}\right)$$

where z_α is the αth quantile from the standard normal distribution (Imbens and Spady 2002).

2.4 Simulation Study

A simulation study is conducted to demonstrate the uses and properties of our GMM estimators of the variance using the adjusted canonical parameter in the quasi-exponential family (McCullagh and Nelder 1989; Dobson and Barnett 2008). In Sections 2.4.1 and 2.4.2 we examine the mean-variance relationship parameters in count and binary data, respectively. Data simulated with sample sizes of 500 are analyzed over 5,000 iterations. The means $\boldsymbol{\mu}$ are obtained through the link function g such that $g(\mu_i) = 2 + 5X_i$, with a log link for the Poisson data and a logit link for the binomial data. The single predictor X comes from the Uniform (0, 1.5) distribution for the Poisson simulation and from the standard normal distribution for the binomial simulation. In Section 2.4.3, we evaluate the impact of sample size on the mean-variance estimates for the binomial distribution. Each sample size simulation was replicated with 100 iterations.

2.4.1 Poisson Distribution

The Poisson distribution has a mean-variance relationship in terms of the canonical parameter as $h'(\theta) = \mu$ and the variance of the Poisson distribution is $var(Y) = \mu$. Thus, from the simulated data we expect the mean-variance parameter estimates $\hat{\psi}$ and $\hat{\lambda}$ to be equal or close to 1 (Table 2.1).

Table 2.1. Poisson Simulation Results

Parameter	**Estimate**	**Standard Error**
ψ	0.996	0.002
λ	0.996	0.021

The results are in agreement for both parameters, thereby supporting the relation in the canonical form. We fix the parameter ψ at 1 and select the form $h'(\theta) = \mu$ as our responses are on the integer scale. The estimate for λ with a fixed value of ψ is given in Table 2. The canonical parameter $h'(\theta)$ is robust to misspecification. The estimate for λ when the canonical parameter is specified as $h'(\theta) = \mu^2$, such that $var(Y) = (\mu^2)^\lambda$, is also provided in Table 2. We obtain a value for $\hat{\lambda}$ close to 0.5, as expected.

Table 2.2. Poisson Simulation Results, $\psi = 1$

$\boldsymbol{h'(\theta)}$	**Estimate**	**Standard Error**
μ	0.996	0.018
μ^2	0.498	0.011

2.4.2 Binomial Distribution

The mean-variance relationship for a binomial random variable Y is evaluated using the canonical parameter $h'(\theta) = p(1-p)$, as $\theta = logit(p)$. The simulation study evaluates the performance of the GMM estimators under the true mean-variance relationship for the binomial distribution with m = 1 and the results support the adequacy of the model, Table 2.3.

Table 2.3. Binomial Simulation Results

	Estimate	**Standard Error**
ψ	1.031	0.349
λ	1.001	0.221

The GMM estimators have values of $\hat{\psi}$ = 1.031 and $\hat{\lambda}$ = 1.001, with standard errors are consistent with adequate performance. The technique accurately obtains the parameter values of ψ and λ as 1. This is seen through the confidence intervals which include the true value.

2.4.3 Sample Size Evaluation

The simulation study examines the performance of the GMM mean-variance estimation for the binomial distribution with $m = 1$ over various sample sizes, n = 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500. The estimates, standard errors, and percent significant (simulations where the estimate was found to be significantly greater than 1) are obtained for ψ and λ over 5000 simulations.

Table 2.4. Sample Size Simulation Results

		Estimate	**Standard Error**	**Percent Significant**
n = 30	ψ	6.308	6.991	14.8%
	λ	1.260	0.813	6.5%
n = 40	ψ	3.977	3.686	11.6%
	λ	1.111	0.764	3.0%
n = 50	ψ	2.623	2.393	9.6%
	λ	1.045	0.692	2.4%
n = 60	ψ	2.114	1.959	9.0%
	λ	1.025	0.640	2.6%
n = 70	ψ	1.771	1.615	7.8%
	λ	1.035	0.598	2.5%
n = 80	ψ	1.741	1.369	7.4%
	λ	1.015	0.556	2.3%
n = 90	ψ	1.416	1.252	6.9%
	λ	1.023	0.530	1.9%
n = 100	ψ	1.435	1.246	6.1%
	λ	1.028	0.506	1.6%
n = 200	ψ	1.114	0.621	3.9%
	λ	1.011	0.353	0.5%
n = 300	ψ	1.069	0.476	2.6%
	λ	1.007	0.288	0.4%
n = 400	ψ	1.050	0.401	2.1%
	λ	1.007	0.249	0.4%
n = 500	ψ	1.035	0.351	2.3%
	λ	1.002	0.222	0.4%

The simulation study suggests that the mean-variance parameter estimates vary more as the sample size decreases. For example, for a sample size of 50, the estimate for the parameter ψ reveals significance in 9.6% of simulated cases. However, for a sample size of 80, the parameter estimates are $\hat{\psi}$ = 1.741 and $\hat{\lambda}$ = 1.015 and the estimate for the parameter ψ is significant in only 7.4% of simulations.

2.5 Numerical Examples

The GMM approach to estimate the variance is useful in many applications, particularly with count and binary outcomes which often exhibit overdispersion. We

revisit the count of snow geese in a flock, which was previously analyzed (Tsou 2011), to compare our approach to the adjusted likelihood approach. We also examine the mean-variance relationship for children's morbidity in the Philippines Bukidnon Province for overdispersion.

2.5.1 Snow Geese Count

We consider two observers estimating the number of snow geese in a flock from an airplane to test aerial survey methods. The observers made estimates of the number of geese, which were compared to the true number of geese determined by a photograph (Weisberg 1985). The observers counted geese from 45 flocks. Figure 2.1 shows one of the observers counts and indicates that there may be overdispersion, as the variance in the count of snow geese increases as the size of the flock increases.

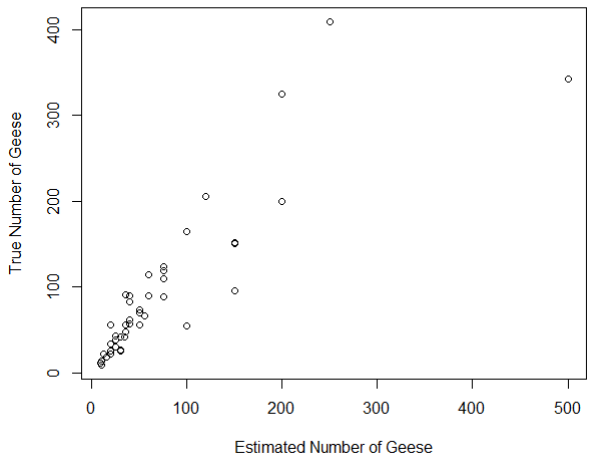


Figure 2.1. Plot of the Estimated Number of Snow Geese Compared to the True Number of Snow Geese in a Flock

To estimate the two parameters in the variance relationship, one may consider three moment conditions thereby resulting in more efficient estimates. We demonstrate

the use of an extra moment condition, Equation 2.3, and estimate $\widehat{var}(Y_i) = 1.178\mu_i^{1.780}$ with confidence intervals of (1.164,1.192) and (1.674,1.887) for ψ and λ, respectively. Tsou (2011) analyzed these data and obtained the mean-variance relationship using a likelihood approach by holding ψ fixed as 0.092 and estimated $\hat{\lambda}$ = 2.027 with a standard error of 0.257. For comparative purposes, we apply the GMM technique with ψ fixed and obtain an estimate of $\hat{\lambda}$ = 2.153 with a standard error of 0.043.

2.5.2 Philippines Morbidity

We evaluate data on children's health collected by the International Food Policy Research Institute in the Philippines (Bhargava 1994). Information including morbidity, age, gender, and body mass index (BMI) are available for 370 children. We consider the first visit for each child to evaluate the variance across children. The binary outcome morbidity indicates whether the child was sick or not. We evaluate overdispersion in the data through estimating the canonical form of the variance $var(y_i) = \psi h'(\theta)^{\lambda}$ for $h'(\theta) = p(1-p)$. The probability of morbidity for each child p_i is estimated using a generalized linear mixed model due to the repeated measurements. Using the GMM approach for estimating the mean-variance relationship, we obtain $\hat{\psi}$ = 0.841 and $\hat{\lambda}$ = 0.891 with standard errors 0.519 and 0.407, respectively. These results indicate that the morbidity data for the first visit of each child are not overdispersed.

2.6 Conclusions

Although it is common to assume that the variance of a random variable is a function of the mean, it is often the case that the true variance in the data may be inflated due to underlying correlation, resulting in overdispersion. Overdispersion is a common phenomenon when analyzing count and binary data, particularly in longitudinal or

clustered data. Using likelihood methods to analyze data relies on the true mean-variance relationship, which assumes that the underlying distribution is correctly specified. The generalized method of moments approach is an alternative technique that produces estimators that do not require distributional assumptions or a pre-assumed mean-variance relationship. It is computationally tractable and has many nice properties. Our simulation study verifies that the GMM estimator produces unbiased and consistent estimates with low standard errors. The performance demonstrates that the GMM estimation technique identifies the mean-variance relationship in the data. The estimators are appropriate and reliable for estimating the variance parameters ψ and λ. The simulation study also demonstrated that the accuracy of the mean-variance relationship depends on sample size, and caution should be used when applying this approach to datasets of small sample sizes. The two numerical examples demonstrate how estimation of both variance parameters can be used to identify overdispersion. Our GMM approach is a comparable alternative to likelihood based methods of estimating the form of the variance and serves as an indicator for overdispersion and potential violations to model assumptions. More importantly, the parameter estimates can be utilized to fit an overdispersed model through a generalized quasi-likelihood model, which takes into account the mean-variance relations in the data (Wedderburn 1974).

CHAPTER 3

ADJUSTED CANONICAL GENERALIZED QUASI-LIKELIHOOD

Abstract

Generalized quasi-likelihood models are useful for analyzing data without an underlying distributional assumption. These models rely on the mean-variance relationship of the data and have many good properties such as unbiased estimates and small standard errors. In correlated data, the underlying mean-variance relationship may be shifted although the distribution is still a member of the quasi-exponential family. We propose fitting generalized quasi-likelihood models to correlated data using alternative estimation of the covariance matrix. We implement the canonical parameter adjustment to the mean-variance relationship estimated using generalized method of moments and extend it for use in hierarchical data. We demonstrate the performance of this adjusted generalized quasi-likelihood approach through a simulation study and apply this modeling technique to Filipino children's morbidity data and adolescent obesity data in the United States.

3.1 Introduction

When analyzing data under particular distributional assumptions, we assume a prescribed mean-variance relationship exists in the data. However, it is often the case that a given dataset will exhibit overdispersion (larger variance than expected) by the underlying distribution. In this case, the true distribution is unknown and corrections for the inflated variance need to be incorporated into the analyses. Ignoring possible overdispersion is a naïve approach to analyze the data and can lead to poor and inaccurate

estimates. Binary and count data commonly exhibit overdispersion, as well as hierarchical data. Longitudinal and clustered data analyses are correlated due to similarities between clusters and within clusters.

Generalized linear models are widely used as a standard tool in regression analysis when the distribution is a member of the exponential family. However, it is well known that traditional generalized linear models cannot be employed to obtain regression parameter estimates when there is correlated data or significant overdispersion. Such data require statistical models such as generalized estimating equations, generalized linear mixed models, and joint modeling of the mean and dispersion (Wilson and Lorenz 2015). Generalized estimating equations account for correlation through the selection of a covariance structure for the correlated responses (Liang and Zeger 1986). Lee and Nelder (2000) presented the use of generalized linear mixed models to model overdispersion in non-normal data. Mixed models incorporate random effects, such as random intercepts and random slopes, to account for correlation due to clustering (Breslow and Clayton 1993). The joint modeling of the mean and the variance uses an additional submodel to address the dispersion parameter in a generalized linear model context (Smyth 1989). The joint modeling of the mean and variance accounts for variation in both the mean and the variance submodels and has been extended to joint modeling in hierarchical generalized linear model structures (Smyth and Verbyla 1999; Lee and Nelder 2006).

In cases where the underlying distribution is unspecified, a quasi-likelihood approach can be implemented (Wedderburn 1974). This modeling technique does not assume an underlying distribution. In quasi-likelihood modeling, the distribution assumption is relaxed through the specification of a variance function. Such approach

requires the specification of the mean-variance relationship and estimates of the covariance matrix in a quadratic form to obtain the regression parameter estimates and corresponding standard errors. The quasi-likelihood approach has many good properties, including unbiased estimates and smaller standard errors as compared to alternative methods (Wang and Wilson 2017).

In this paper, a generalized quasi-likelihood (GQL) model to fit correlated data is presented which uses an adjusted canonical parameterization of the mean-variance relationship in the covariance. The adjusted canonical parameter provides a generalization that makes this approach feasible for all models with unknown distributions but is believed to be a member of the quasi-exponential family. The adjusted GQL model negates the need for distributional assumptions as required with a maximum likelihood estimation approach, and incorporates the empirical variance in the data. In Section 3.2, we review a generalized quasi-likelihood approach that estimates the regression parameters and the variance components in a clustered data setting (Sutradhar 2004). In Section 3.3, we propose an alternative model to analyze data using generalized quasi-likelihood while accounting for the overdispersion. The method is simple as it incorporates an adjustment to the canonical parameter. In Section 3.4, we validate the performance of the adjusted generalized quasi-likelihood model with GMM estimates of the mean-variance relation through a simulation study. In Section 3.5, the GQL model with adjusted canonical parameters are used to analyze the children's health in the Philippines (Bhargava 1994) and obesity data collected through the National Longitudinal Study of Adolescent to Adult Health (Add Health) (Harris, et al. 2009).

3.2 Generalized Quasi-likelihood Models

Consider n vectors of observations $\boldsymbol{y_1}, \dots, \boldsymbol{y_n}$ where $\boldsymbol{y_i} = \left(y_{i1}, \dots . y_{in_i}\right)$ for $i = 1, \dots n$ where y_{ij} follows a distribution from the exponential family with link function $g(\cdot)$ such that $g\left(\mu_{ij}\right) = \eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \alpha_i$ for k covariates where $\alpha_i \sim N(0, \sigma^2)$ is a random intercept for cluster i. The linear component can also be written as $\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \sigma\xi_i$ where $\xi_i \sim N(0,1)$ as $\xi_i = \alpha_i/\sigma$. We estimate $\boldsymbol{\beta}$ and σ using GQL. Let the response vector be $\boldsymbol{S_i} = (y'_i, u'_i)$ where $\boldsymbol{y'_i} = (y_{i1}, \dots, y_{in_i})$ and $\boldsymbol{u'_i} = (\boldsymbol{u'_{i1}}, \boldsymbol{u'_{i2}})$ contains the pairwise products where $\boldsymbol{u_{i1}} = \left(y_{i1}^2, \dots, y_{in_i}^2\right)$ and $\boldsymbol{u_{i2}} = (y_{i1}y_{i2}, \dots, y_{ij}y_{ij'}, \dots, y_{i(n_i-1)}y_{in_i})$. The vector of model parameters is $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma)'$ and $\boldsymbol{M_i}(\theta)$ is the mean of the response vector $\boldsymbol{S_i}$. Let $\boldsymbol{\Omega_i}(\theta)$ be the covariance matrix for $\boldsymbol{S_i}$ as $\boldsymbol{\Omega_i}(\theta) = \left(\omega_{ij}\right)$. Then, the generalized quasi-likelihood estimating equation

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{M'_i}(\theta)}{\partial \theta} \boldsymbol{\Omega}_i^{-1}(\theta)[\boldsymbol{S_i} - \boldsymbol{M_i}(\theta)] = 0 \qquad (3.1)$$

is used to obtain the GQL estimates of $\boldsymbol{\beta}$ and σ (Wedderburn 1974; Sutradhar 2004). The rth finite moments of y_{ij} determine the GQL estimates, denoted using the function $g_{(r)}(\eta_{ij})$.

The mean of the response vector, $\boldsymbol{M_i}(\theta)$, is evaluated based on

$$\boldsymbol{M_i}(\theta) = E(\boldsymbol{S_i}) = E\left(Y_{i1}, \dots, Y_{in_i}, Y_{i1}^2, \dots, Y_{in_i}^2, Y_{i1}Y_{i2}, \dots, Y_{i(n_i-1)}Y_{in_i}\right)$$

where

$$E\left(Y_{ij}\right) = \mu_{ij}(\theta) = E\left[g_{(1)}\left(x'_{ij}\beta + \sigma\xi\right)\right]$$

$$E\left(Y_{ij}^2\right) = m_{ijj}(\theta) = E\left[g_{(2)}\left(x'_{ij}\beta + \sigma\xi\right)\right]$$

$$E\left(Y_{ij}Y_{ik}\right) = m_{ijk}(\theta) = E\left[g_{(1)}\left(x'_{ij}\beta + \sigma\xi\right)g_{(1)}(x'_{ik}\beta + \sigma\xi)\right].$$

The partial derivative matrix $\frac{\partial \boldsymbol{M}_i'(\theta)}{\partial \theta}$ has dimension $(p+1) \times \{n_i(n_i+1)/2\}$, where the partial derivatives are

$$\frac{\partial \mu_{ij}(\theta)}{\partial \beta} = E\left[\tilde{g}_{(1)}\left(x_{ij}'\beta + \sigma\xi\right)\right]x_{ij}'$$

$$\frac{\partial m_{ijj}(\theta)}{\partial \beta} = E\left[\tilde{g}_{(2)}\left(x_{ij}'\beta + \sigma\xi\right)\right]x_{ij}'$$

$$\frac{\partial m_{ijk}(\theta)}{\partial \beta} = E\left[\tilde{g}_{(1)}\left(x_{ij}'\beta + \sigma\xi\right)g_{(1)}(x_{ik}'\beta + \sigma\xi)\right]x_{ij}' + E\left[g_{(1)}\left(x_{ij}'\beta + \sigma\xi\right)\tilde{g}_{(1)}(x_{ik}'\beta + \sigma\xi)\right]x_{ik}'$$

$$\frac{\partial \mu_{ij}(\theta)}{\partial \sigma} = E\left[\xi\tilde{g}_{(1)}\left(x_{ij}'\beta + \sigma\xi\right)\right]x_{ij}'$$

$$\frac{\partial m_{ijj}(\theta)}{\partial \sigma} = E\left[\xi\tilde{g}_{(2)}\left(x_{ij}'\beta + \sigma\xi\right)\right]x_{ij}'$$

$$\frac{\partial m_{ijk}(\theta)}{\partial \sigma} = E\left[\xi\left\{\tilde{g}_{(1)}\left(x_{ij}'\beta + \sigma\xi\right)g_{(1)}(x_{ik}'\beta + \sigma\xi) + g_{(1)}\left(x_{ij}'\beta + \sigma\xi\right)\tilde{g}_{(1)}(x_{ik}'\beta + \sigma\xi)\right\}\right]$$

where $\tilde{g}_{(r)}(\cdot)$ is the first derivative of $g_{(r)}(\cdot)$. The covariance matrix is

$$\boldsymbol{\Omega_i} = \begin{bmatrix} \boldsymbol{\Sigma_i} & \boldsymbol{P_i} \\ \boldsymbol{P_i'} & \boldsymbol{Q_i} \end{bmatrix}$$

where $\boldsymbol{\Sigma_i} = cov(Y_i)$, $\boldsymbol{P_i} = cov(Y_i, U_i')$ and $\boldsymbol{Q_i} = cov(U_i)$. The diagonal elements of $\boldsymbol{\Sigma_i}$, the variances of Y_i, are

$$\sigma_{ijj} = Var\left(Y_{ij}\right) = m_{ijj}(\theta) - \mu_{ij}^2(\theta).$$

with off diagonal elements

$$\sigma_{ijk} = Cov\left(Y_{ij}, Y_{ik}\right) = m_{ijk}(\theta) - \mu_{ij}(\theta)\mu_{ik}(\theta).$$

The matrix $\boldsymbol{P_i}$ is of dimension $n_i \times \{n_i(n_i+1)/2\}$ and contains $cov(Y_{ij}, Y_{ij}^2)$, $cov(Y_{ij}, Y_{ij}Y_{il})$ and $cov(Y_{ij}, Y_{ik}Y_{il})$.

For $j = k = l$, $cov\left(Y_{ij}, Y_{ij}^2\right) = p_{ijjj}(\theta) = E\left[g_{(3)}\left(x_{ij}'\beta + \sigma\xi\right)\right] - \mu_{ij}(\theta)m_{ijj}(\theta)$.

For $j = k \neq l$ and $j = l \neq k$, the covariance elements are

$cov\left(Y_{ij}, Y_{ij}Y_{il}\right) = E\left(Y_{ij}^2 Y_{il}\right) - \mu_{ij}(\theta)m_{ijl}(\theta) = E\left[g_{(2)}\left(x_{ij}'\beta + \sigma\xi\right)g_{(1)}(x_{il}'\beta + \sigma\xi)\right] - \mu_{ij}(\theta)m_{ijl}(\theta)$.

For $j \neq k \neq l$, $cov\left(Y_{ij}, Y_{ik}Y_{il}\right) = p_{ijkl}(\theta) = E\left[g_{(1)}\left(x_{ij}'\beta + \sigma\xi\right)g_{(1)}(x_{ik}'\beta + \sigma\xi)g_{(1)}(x_{il}'\beta + \sigma\xi)\right] - \mu_{ij}(\theta)m_{ikl}(\theta)$.

In the covariance matrix, $\boldsymbol{Q_i}$ contains $cov(Y_{ij}Y_{ik}, Y_{il}Y_{iw})$ with dimension $\{n_i(n_i + 1)/2\} \times \{n_i(n_i + 1)/2\}$.

For $j = k = l = m$, $cov\left(Y_{ij}^2, Y_{ij}^2\right) = q_{ijjjj}(\theta) = E\left[g_{(4)}\left(x_{ij}'\beta + \sigma\xi\right)\right] - m_{ijj}^2(\theta)$.

For $j = k \neq l \neq w$, $cov\left(Y_{ij}^2, Y_{il}Y_{iw}\right) = q_{ijjlw}(\theta) = E\left[g_{(2)}\left(x_{ij}'\beta + \sigma\xi\right)g_{(1)}(x_{il}'\beta + \sigma\xi)g_{(1)}(x_{iw}'\beta + \sigma\xi)\right] - m_{ijj}(\theta)m_{ilw}(\theta)$.

For $j \neq k \neq l \neq w$,

$cov\left(Y_{ij}Y_{ik}, Y_{il}Y_{iw}\right) = q_{ijklw} = E\left[g_{(1)}\left(x_{ij}'\beta + \sigma\xi\right)g_{(1)}(x_{ik}'\beta + \sigma\xi)g_{(1)}(x_{il}'\beta + \sigma\xi)g_{(1)}(x_{iw}'\beta + \sigma\xi)\right] - m_{ijk}(\theta)m_{ilw}(\theta)$.

Then, the quasi-likelihood estimate $\widehat{\boldsymbol{\theta}}_{QL} = \left(\widehat{\boldsymbol{\beta}}_{QL}', \hat{\sigma}_{QL}\right)'$ is found using Newton-Raphson iteration as

$$\widehat{\boldsymbol{\theta}}_{QL}(t+1) = \widehat{\boldsymbol{\theta}}_{QL}(t) + \left[\sum_{i=1}^{n} \frac{\partial \boldsymbol{M}_i'(\theta)}{\partial \theta} \boldsymbol{\Omega}_i^{-1} \frac{\partial \boldsymbol{M}_i(\theta)}{\partial \theta}\right]_{(t)}^{-1} \left[\sum_{i=1}^{n} \frac{\partial \boldsymbol{M}_i'(\theta)}{\partial \theta} \boldsymbol{\Omega_i}^{-1}(\theta)[\boldsymbol{S_i} - \boldsymbol{M_i}(\theta)]\right].$$

The covariance of the quasi-likelihood estimator is

$$\hat{V}\left(\widehat{\boldsymbol{\theta}}_{QL}\right) = \left[\sum_{i=1}^{n} \frac{\partial \boldsymbol{M}_i'(\theta)}{\partial \theta} \boldsymbol{\Omega}_i^{-1} \frac{\partial \boldsymbol{M}_i(\theta)}{\partial \theta}\right]^{-1}.$$

GQL estimators are consistent and efficient (Sutradhar 2004). Specification of the GQL model is important as consistency of the regression parameter estimates depends on

correctly specifying link function and efficiency depends on a correctly specified variance function.

3.3 Adjusted Quasi-likelihood Models with Canonical Parameterization

We postulate that the variance parameter estimates $\hat{\psi}$ and $\hat{\lambda}$ obtained through GMM estimation (Chapter 2) is applicable to modeling the mean-variance relationship in overdispersed data and to estimate the covariance matrix for inference. In this section, we use quasi-likelihood estimation in a semiparametric model with a correlated structure based on the canonical parameter representation in the mean-variance relationship. It is natural to find some deviation from the relation between the mean and variance when dealing with correlated data. As such, we rely on the first two moments of the response based on knowledge of the scale of the responses.

3.3.1 Mean-Variance Estimation in Hierarchical Data

In Chapter 2, the mean-variance relation was presented as a useful method for identifying deviations from the assumed variance. The canonical mean-variance form extends the estimation to distributions in the exponential family and estimates the true mean-variance relation for moderate sample sizes. Initially, the mean-variance estimation method was used with cross-sectional data. We extend the estimation procedure to identify the mean-variance relation in clustered data and use the estimates in generalized quasi-likelihood modeling.

Let the random variables y_{ij}, for the jth observed value in cluster i, $j = 1, \dots, n_i$ and $i = 1, \dots, n$, have means μ_{ij} related to k covariates and random effect α_i through the link function $g(\cdot)$ such that $E(Y_{ij}) = \mu_{ij}$ and $g(\mu_{ij}) = \eta_{ij} = x'_{ijk}\beta + \alpha_i$ where the random effect α_i represents the variation between clusters such that $\alpha_i \sim N(0, \sigma^2)$.

Further, consider $\xi_i = \alpha_i/\sigma$, where $\xi_i \sim N(0,1)$ so the linear predictor becomes $g(\mu_{ij}) = x'_{ijk}\beta + \sigma\xi_i$.

Recall the general mean-variance relationship, $var(Y) = \psi[h'(\theta)]^{\lambda}$, where θ is the canonical parameter and $h'(\theta)$ is the first derivative of the inverse canonical link. For cross sectional data, the variance parameters ψ and λ were estimated from all the data based on generalized method of moments, (2.2a) and (2.2b). In clustered data, there is added correlation that must be addressed in any models. For two-level hierarchical, consider n clusters where n is of moderate size. Let an observation be sampled from each cluster denoted as $y_{1(j_1)}, y_{2(j_2)} \dots, y_{n(j_n)}$ where (j_i) is the sampled observation number for the ith group where $1 \leq j \leq n_i$. We apply the GMM procedure to the observations of sample size n_i to estimate the mean-variance relationship parameters ψ and λ for the sampled data. The estimated variances are

$$\widehat{var}(Y_{ij}) = \hat{\psi}[h'(\hat{\theta}_{ij})]^{\hat{\lambda}}.$$

We select m random samples of observations from each cluster and calculate the mean-variance parameters in each sample. Thus, the average of $\hat{\psi}_k$ and $\hat{\lambda}_k$ for $k = 1, \dots, m$ provide parameter estimates. For the situation where there are few clusters and n_i is of moderate size, the mean-variance relation parameters can be estimated using a moderate number of observations within each cluster.

3.3.2 Adjusted Generalized Quasi-likelihood

From use the generalized quasi-likelihood estimating equation (3.1), with $\boldsymbol{M_i}$, $\frac{\partial \boldsymbol{M}'_i(\theta)}{\partial \theta}$, and the covariance matrix $\boldsymbol{\Omega_i}(\theta)$, we estimate the regression parameters $\boldsymbol{\beta}$ and the variance of the random effect σ. While GQL is known to perform well and produce

consistent and efficient estimators (Sutradhar 2004), it depends on the estimate of the variance of S_i, which we obtain from the canonical mean-variance relationship form $\psi\left[h'\left(\theta_{ij}\right)\right]^{\lambda}$. The covariance matrix $\boldsymbol{\Omega}_i(\boldsymbol{\beta}, \sigma, \psi, \lambda\ )$ is estimated with diagonal elements in $\boldsymbol{\Sigma_i}$ of $\sigma_{ii} = \sigma_{ii}(\beta, \sigma, \psi, \lambda)$ where

$$\sigma_{ijj} = var\left(Y_{ij}\right) = \psi\left[h'\left(\theta_{ij}\right)\right]^{\lambda}$$

for estimates of ψ and λ estimated using GMM on a random sample of an observation from each cluster. For given ψ and λ, the GQL estimates of $\boldsymbol{\beta}$ and σ from (3.1) are unbiased. The GQL estimators are consistent and efficient as μ_{ij} is the mean of y_{ij} and the weight matrix $\boldsymbol{\Sigma_i}$ reflects the estimated covariance.

3.4 Simulation Study

Consider simulated binary response data in a two-level hierarchical data structure. Each dataset contains 100 clusters, each has 5 observations in each cluster (for a total of 500 observations). The random intercept α_i associated with each cluster is generated from $N(0, \sigma_\alpha^2)$ with σ_α = 1, 2, 3, 4, or 5. The linear predictor is $\eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \alpha_i$, where $\beta_1 = \beta_2 = 1$ and X_1 and X_2 are generated from standard normal distributions. We fit an adjusted GQL model (averaged across ten random samples), GQL model, and a generalized linear mixed model (GLMM) to the simulated data for 500 iterations, Table 3.1.

Table 3.1. GQL Simulation Results for Binary Data

		$\widehat{\boldsymbol{\beta}}_1$ **(SE)**	$\widehat{\boldsymbol{\beta}}_2$ **(SE)**	$\widehat{\boldsymbol{\sigma}}$ **(SE)**
$\sigma = 1$	Adjusted GQL	1.0071 (0.1523)	0.9439 (0.1547)	1.1602 (0.2555)
	GQL	1.0039 (0.1417)	0.9982 (0.1413)	1.0022 (0.2061)
	GLMM	0.9988 (0.1369)	1.0135 (0.1376)	0.9557 (-)
$\sigma = 2$	Adjusted GQL	1.0142 (0.1879)	1.0272 (0.1890)	2.0553 (0.3858)
	GQL	1.0095 (0.1635)	1.0224 (0.1642)	2.0441 (0.2987)
	GLMM	1.0021 (0.1619)	1.0147 (0.1627)	1.9456 (-)
$\sigma = 3$	Adjusted GQL	1.0264 (0.2372)	1.0357 (0.2384)	3.1000 (0.6676)
	GQL	1.0137 (0.1860)	1.0279 (0.1868)	3.0558 (0.4511)
	GLMM	0.9903 (0.1814)	1.0043 (0.1820)	2.8352 (-)
$\sigma = 4$	Adjusted GQL	1.0503 (0.2918)	1.0663 (0.2945)	4.2475 (1.0827)
	GQL	1.0230 (0.2089)	1.0354 (0.2097)	4.0954 (0.6486)
	GLMM	0.9734 (0.1987)	0.9853 (0.1990)	3.6486 (-)
$\sigma = 5$	Adjusted GQL	1.0528 (0.3367)	1.0742 (0.3397)	5.3381 (1.5583)
	GQL	1.0161 (0.2291)	1.0332 (0.2302)	5.1082 (0.8747)
	GLMM	0.9364 (0.2074)	0.9521 (0.2080)	4.3366 (-)

The adjusted GQL model produces accurate estimates of the regression parameters β_1, β_2, and random effect variance. The parameter estimates are similar across the three methods (mixed, GQL, and adjusted GQL). The simulation results

suggest that the adjusted GQL model recovers the true values when relying on the estimated mean-variance relationship in the covariance matrix.

3.5 Numerical Examples

We analyze two longitudinal examples, children's morbidity in the Philippines Bukidnon Province and adolescent obesity data in the United States. We fit an adjusted generalized quasi-likelihood model using the canonical mean-variance relationship for the covariance structure. Parameter estimates are compared to the generalized linear mixed model and generalized quasi-likelihood models.

3.5.1 Philippines Children's Morbidity Study

Data on children's health were obtained by the International Food Policy Research Institute in the Philippines (Bhargava 1994). Information pertaining to morbidity including age in months and body mass index (BMI) are available for 370 children, collected over three visits. The binary outcome morbidity indicates whether the child was sick at the time of the visit. We address potential overdispersion in the data through the canonical form of the variance $var(y_{ij}) = \psi h'(\theta_{ij})^{\lambda}$ where $h'(\theta_{ij}) = p_{ij}(1 - p_{ij})$. The mean-variance parameter estimates, obtained from 100 random samples of one observation per child, are $\hat{\psi}$ = 1.16 and $\hat{\lambda}$ = 1.07 with standard errors 0.97 and 0.53, respectively. These results lead us to believe the data are not overdispersed (test statistics Z_{ψ} = 0.16 and Z_{λ} = 0.13), however, we estimate the regression parameters for age and BMI, and the standard deviation of the random effect for children using adjusted GQL, GQL, and GLMM for comparison.

Table 3.2. Children's Morbidity Model Estimates and Standard Errors

		$\boldsymbol{\beta_{Age}}$	$\boldsymbol{\beta_{BMI}}$	$\boldsymbol{\sigma}$
Adjusted GQL	Estimate	-0.0187	-0.0175	0.8085
	Std. Error	0.0047	0.0136	0.1572
GQL	Estimate	-0.0185	-0.0184	0.8114
	Std. Error	0.0047	0.0135	0.1529
GLMM	Estimate	-0.0183	-0.0178	0.7199
	Std. Error	0.0045	0.0130	-

The model estimates using the three approaches are similar, Table 3.2. The estimate of σ is 0.7199 for the generalized linear mixed model, 0.8114 for the generalized quasi-likelihood approach, and 0.8085 for the adjusted generalized quasi-likelihood approach. While the estimate of σ using the adjusted GQL approach is slightly smaller than the estimate from GQL, they do not differ significantly which is expected as the mean-variance parameters did not indicate significant overdispersion.

3.5.2 Add Health Obesity Study

The Add Health Study is a longitudinal study in the United States of adolescents in 7th through 12th grade, with information collected over four waves of interviews between 1994 and 2008 (Harris, et al. 2009). The factors associated with obesity for 2712 adolescents include activity scale and feeling scale, ratings of physical activity and emotional health. Obesity is binary, indicating whether the adolescent was obese at the time of the interview. We use 10 random samples to obtain the mean-variance parameter estimates $\hat{\psi}$ = 0.71 and $\hat{\lambda}$ = 0.82 with standard errors 0.10 and 0.08, respectively. The result indicates that significant deviation from the assumed mean-variance relationship

(test statistics Z_ψ = -3.02 and Z_λ = -2.09). Thus, making use of the true mean-variance form improves the model fit. The model parameter estimates and standard errors are provided, Table 3.3.

Table 3.3. Adolescent Obesity Model Estimates and Standard Errors

		$\boldsymbol{\beta_{Activity\ Scale}}$	$\boldsymbol{\beta_{Feeling\ Scale}}$	$\boldsymbol{\sigma}$
Adjusted GQL	Estimate	-1.2078	-0.5623	2.1885
	Std. Error	0.0515	0.0698	0.0978
GQL	Estimate	-1.1006	-0.5518	1.9993
	Std. Error	0.0442	0.0687	0.0898
GLMM	Estimate	-1.3438	-0.6527	2.6900
	Std. Error	0.0472	0.0726	-

The regression parameter estimates for activity scale and feeling scale are similar. However, the estimates of the standard deviation of the random effect σ vary between the three models. In the generalized linear mixed model and GQL model, the estimates are $\hat{\sigma}_{GLMM}$ = 2.690 and $\hat{\sigma}_{GQL}$ = 1.999. For the adjusted GQL model, the estimate $\hat{\sigma}_{AdjGQL}$ = 2.189 is outside the confidence interval for $\hat{\sigma}_{GQL}$. This significant difference in the estimators is realized due to the incorporation of the true variance in the estimation procedure.

3.6 Conclusions

Correlation on account of clustering can alter the variance in a dataset. This shift in the mean-variance relation due to correlation is addressed with additional modeling. While generalized linear mixed models and generalized quasi-likelihood models account for correlation through the variance of the random effect, one can also address the correlation through a mean-variance relation. The alternative parameterization presented

uses the covariance matrix in GQL to model extra variation. The simulation study demonstrates the performance of incorporating the canonical mean-variance relationship in the GQL covariance matrix. The numerical examples, Philippines morbidity and Add Health, demonstrate that one can rely on the estimated mean-variance relationship in the presence of overdispersion. The adjusted GQL model makes use of the mean-variance relationship to address correlation in the data. The flexibility in the canonical mean-variance approximation and modeling through GQL makes this model appropriate for any distribution in the quasi-exponential family.

CHAPTER 4

JOINT MODELING OF MEAN, VARIANCE, SKEWNESS, AND KURTOSIS

Abstract

When modeling data, it is common to model the mean, or the first moment of the distribution of the responses, and determine which covariates influence the outcome. One often improves the model fit by evaluating and accounting for the unexplained variation. Such is the case when we jointly model the mean and the variance. This paper expands on this notion and introduces the joint modeling of the mean (first moment), variance (second moment), skewness (third moment) and kurtosis (fourth moment) to provide an improved fit to the data. For most distributions in the exponential family, the mean, variance, skewness, and kurtosis are related, so one would expect that a covariate that impacts the mean would also influence higher order moments. We use the relationship between the moments to obtain an improved fit. The additional modeling of the skewness and kurtosis is used to trim the data and improve the parameter estimates in the mean and variance submodels. A simulation study demonstrates the performance of joint modeling of the mean, variance, skewness, and kurtosis. We examine proteomic assays for breast cancer screening and identify biomarkers that are associated with a diagnosis of breast cancer. We find that the variance, skewness, and kurtosis are often impacted by the same predictors when the distribution is realized.

4.1 Introduction

When fitting models to data, the aim is usually to understand the relationship between the mean of the outcomes, $E(Y|X)$, and the covariates X. A significant fit allows one to make conclusions about the mean based on the values of the covariates. In the analysis of the mean, one often begins with some knowledge regarding the distribution of $(Y|X)$. Although there are times when the underlying distribution is completely unknown, we can begin with a relationship between the first two moments. When the distribution is known, we use methods such as likelihood or quasi-likelihood to obtain the parameter estimates for the mean of the distribution of responses. In order to improve the model fit, it is natural to consider and model the unexplained variation through a function of the observed y and the predicted mean $\hat{\mu}_{Y|X}$. Smyth (1989) considered the joint modeling of the mean and dispersion by using the unexplained deviation in the mean submodel to improve the model fit. The additional submodel allows one to model how the spread of the data from the mean relates to certain covariates. The covariates in the mean submodel may be included in the dispersion submodel, or additional covariates may be considered.

While the joint modeling of the mean and dispersion submodels accounts for unequal variance across subpopulations, it does not account for higher order moments which are often related to the mean and variance through certain parameters. This paper extends the joint modeling approach to model functions of the skewness and the kurtosis of the distribution. Thus, we add a third and a fourth submodel. These additional measures, relating to the third and fourth moments of the distribution, represent the shape characteristics. Some researchers have considered various estimators for higher order

moments (Joanes and Gill 1998; Jiang 2003), and found that testing for skewness and kurtosis have led to improvements in modeling time series and longitudinal data (Bai and Ng 2005; Soberón and Stute 2017). Similar to location and scale, the measures of skewness and kurtosis help model certain aspects of the distribution of responses. Therefore, it is conceivable to extend joint modeling to incorporate skewness and kurtosis (Balanda and MacGillivray 1988).

This paper introduces the fit of four separate models that are related through functions based on the deviation between the data Y and the predicted conditional mean $\hat{\mu}_{Y|X}$ obtained from the mean submodel. Let γ_1 and γ_2 be the skewness and kurtosis, respectively, defined as

$$\gamma_1 = \frac{E(Y-\mu)^3}{(\sigma^2)^{3/2}}$$

and

$$\gamma_2 = \frac{E(Y-\mu)^4}{(\sigma^2)^2} - 3.$$

where μ is the mean of Y and σ^2 is the variance of Y. The term γ_2 is also referred to as the excess kurtosis since it is relative to the normal distribution, but henceforth we will refer to this simply as the kurtosis. The joint estimation of the mean, variance, skewness, and kurtosis parameters is useful for identifying deviations in the expected skewness and kurtosis under the assumed distribution. Neykov, Filzmoser, and Neytchev (2012) showed that joint modeling of mean and dispersion can be sensitive to outliers in the data. They provided a trimming approach to obtain robust estimators. We also propose a cutoff for trimming based on the skewness and kurtosis to improve the model fit and increases the estimation accuracy for both the mean and the dispersion models.

We use the joint modeling of the mean, variance, skewness, and kurtosis to study the biomarkers the diagnosis of breast cancer. Breast cancer screening methods have rapidly improved with the implementation of proteomic assays. While the assays have increased diagnosis accuracy, researchers are still investigating the relationship between particular biomarkers and breast cancer. Our analysis identifies certain biomarkers associated with a positively skewed probability of having breast cancer, which is a new contribution to the field. The joint modeling of the mean, variance, skewness and kurtosis gives a better understanding of the impact of the covariates on the outcome. In Section 4.2, we review distributional moments, including a characterization of the skewness and the kurtosis through the third and fourth moments, and present a review of joint modeling of the mean and dispersion (Smyth 1989). Section 4.3 introduces the joint modeling of the mean, variance, skewness and kurtosis and the estimation of the regression parameters for each submodel. In addition, we provide a cutoff based on the skewness and kurtosis to trim the data and improve the model fit. In Section 4.4, we conduct a simulation study to demonstrate the advantages of joint modeling of mean, variance, skewness, and kurtosis as opposed to the joint modeling of the mean and dispersion. In Section 4.5, we analyze the breast cancer proteomic assay data and provide some overall conclusions in Section 4.6.

4.2 Background

4.2.1 Notation

Let the random variable Y with mean μ and variance σ^2 be a member of the exponential family which has the form

$$f(y;\theta,\phi) = exp\left\{\frac{(y\theta - b(\theta))}{a(\phi)} + c(y,\phi)\right\} \tag{4.1}$$

where θ is the canonical parameter, ϕ is the dispersion parameter, and the functions a, b, and c are known. The mean of Y is $\mu = E(Y) = b'(\theta)$ and the variance is $\sigma^2 = var(Y) = a(\phi)b''(\theta)$. Thus, the mean-variance relationship is expressed as

$$var(Y) = \frac{\phi}{w}V(\mu)$$

for $a(\phi) = \frac{\phi}{w}$ and $V(\mu)$ as a function representing the relationship between the mean and variance (McCullagh and Nelder 1989).

4.2.2 Skewness and Kurtosis

The skewness γ_1 and the kurtosis γ_2 are related to the third and fourth moments of the distribution and describe its shape through the tails. Both measures are related to the mean (μ) and variance (σ^2) as the skewness,

$$\gamma_1 = \frac{E(Y-\mu)^3}{(\sigma^2)^{3/2}} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\{a(\phi)\}^2 b^{(3)}(\theta)}{\{a(\phi)b''(\theta)\}^{3/2}}$$

is the standardized third central moment and the kurtosis,

$$\gamma_2 = \frac{E(Y-\mu)^4}{(\sigma^2)^2} - 3 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{\{a(\phi)\}^3 b^{(4)}(\theta) + 3\{a(\phi)b''(\theta)\}^2}{\{a(\phi)b''(\theta)\}^2} - 3$$

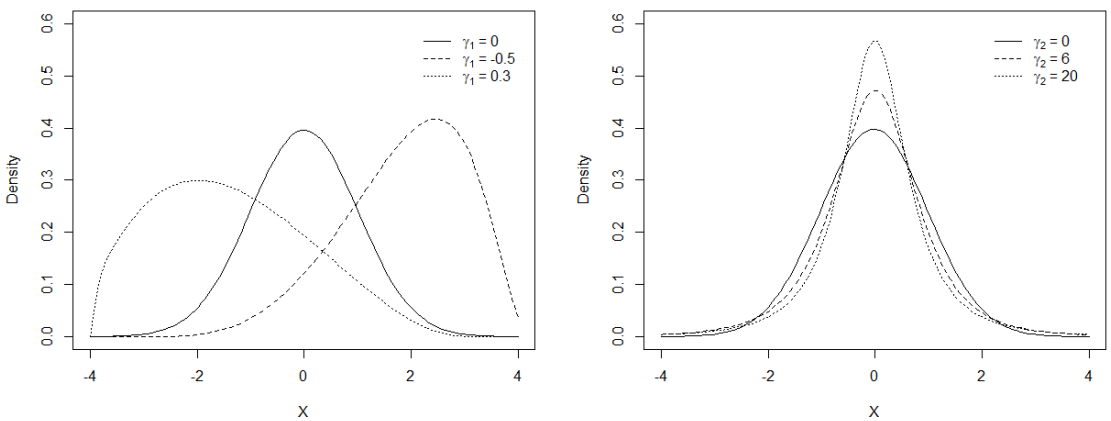
$$= \frac{\{a(\phi)\}^3 b^{(4)}(\theta)}{\{a(\phi)b''(\theta)\}^2}$$

is the standardized fourth central moment less 3, where μ_i is the ith central moment of the distribution. This relationship holds for parameters in the binomial, gamma, normal, and Poisson distributions. Although the normal distribution has a skewness and kurtosis of 0 and no relation between the moments, the binomial, gamma, and Poisson distributions have skewness and kurtosis values that are related to their parameters, Table 4.1.

Table 4.1. Mean, Variance, Skewness, and Kurtosis for Selected Distributions

Distribution	**Mean**	**Variance**	**Skewness**	**Kurtosis**
Binomial(n, p)	np	$np(1-p)$	$(1-2p)/\sqrt{np(1-p)}$	$(6p^2-6p+1)\{np(1-p)\}^{-1}$
Gamma$(\phi^{-1}, \phi\mu)$	μ	$\phi\mu^2$	$2\phi^{1/2}$	6ϕ
Normal(μ, σ^2)	μ	σ^2	0	0
Poisson (λ)	λ	λ	$\lambda^{-1/2}$	λ^{-1}

Skewness and kurtosis are referred to as shape statistics as they indicate the shape of the distribution. Both measures describe the tails of the distribution, as shown in Figure 4.1.



(a) (b)

Figure 4.1. Probability Density Plots for a) Skewness and b) Kurtosis

The skewness measures symmetry in a distribution and describes the relative size of the two tails. A distribution that is completely symmetric, such as the normal distribution, has a skewness of 0. Skewness less than or greater than 0 is related to extreme observations in either the left or the right tails. Distributions that are left skewed, with a left leading tail, are referred to as negatively skewed while distributions that are

right skewed are positively skewed (Fogler and Radcliffe 1974; Groeneveld and Meeden 1984).

The kurtosis is a measure of the combined weight of the tails relative to the rest of the distribution. A large positive kurtosis indicates that the data are greatly concentrated near the mean and declines rapidly from the center with heavy tails on both sides of the mean (Groeneveld and Meeden 1984; DeCarlo 1997). The measure is reported relative to the normal distribution, which has a kurtosis of 0. Distributions are often described as leptokurtic, platykurtic, or mesokurtic, based on the value of the kurtosis. Leptokurtic indicates that the distribution has a kurtosis larger than the normal distribution ($\gamma_2 > 0$) and thus has heavier tails compared to the normal distribution. Platykurtic indicates that the distribution is flat ($\gamma_2 < 0$) and has lighter tails than the normal distribution (Chissom 1970). Mesokurtic distributions have a kurtosis close to 0, and includes the normal distribution. The kurtosis ranges from -2 (a binomial random variable with two equally likely outcomes) to ∞ (a t-distribution with 4 degrees of freedom). Kurtosis can differentiate between two distributions with the same mean and variance, such as a normal distribution with mean 0 and variance 5/3 ($\gamma_2 = 0$) and a t-distribution with 5 degrees of freedom ($\gamma_2 = 6$). While the first two moments of both distributions are the same, the shapes are different as the t-distribution has heavier tails.

Numerous studies have investigated the limiting distributions of the skewness and kurtosis. For an independent normally distributed random variable with n observations, the limiting distributions are $\sqrt{n}\widehat{\gamma_1} \rightarrow N(0, 6)$ and $\sqrt{n}\hat{\gamma}_2 \rightarrow N(0, 24)$ (Kendall and Stuart 1969), where $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are estimators of γ_1 and γ_2. The skewness and kurtosis are often

utilized in tests of normality and studies have indicated that $n\left(\frac{\hat{\gamma}_1^2}{6}+\frac{\hat{\gamma}_2^2}{24}\right)$ converges in distribution to a χ^2 with 2 degrees of freedom (Jarque and Bera 1980; Bera and Jarque 1981). However, when the data are weakly correlated, Bai and Ng (2005) showed

$$\left(\frac{\sqrt{n}\hat{\gamma}_1}{s(\hat{\gamma}_1)}\right)^2+\left(\frac{\sqrt{n}(\hat{\gamma}_2-\gamma_2)}{s(\hat{\gamma}_2)}\right)^2 \xrightarrow{d} \chi_2^2$$

where $s(\hat{\gamma}_1)$ and $s(\hat{\gamma}_2)$ are the asymptotic standard errors of the skewness and kurtosis, respectively. Bai and Ng (2005) also showed that there is bias in the sample kurtosis, particularly in the presence of serial correlation. The estimates of skewness and kurtosis are affected by sample size and a large number of observations is necessary to obtain reliable estimates.

In generalized linear models, the model fit assumes that the higher order moments of the distributions behave independently of lower order moments. For example, in simple linear regression, each observation is assumed to come from a normally distributed population with mean μ, variance σ^2, skewness $\gamma_1 = 0$, and kurtosis $\gamma_2 = 0$, denoted as $Y \sim N(\mu, \sigma^2, \gamma_1, \gamma_2)$. When modeling the mean of this distribution, we see a relationship between the predictor (X) and the outcome. We assume that X does not impact the variance, skewness, or kurtosis, when in fact it may be directly impacting higher order moments. In Figure 4.2, we have the distribution of $Y|X$ as X changes. We assume a normal distribution exists for $Y|X$ at each value (subpopulation) of X, with constant variance and the skewness and kurtosis equal to 0.

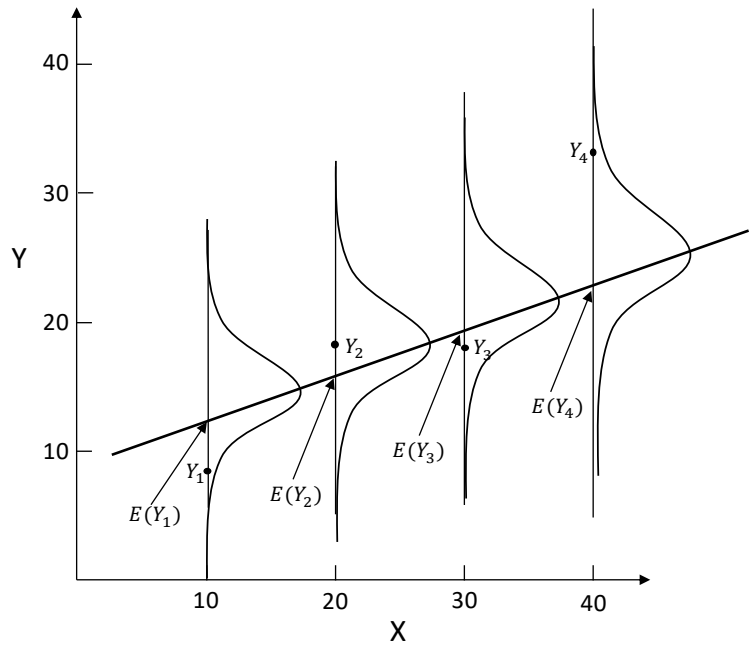


Figure 4.2. Example of Simple Linear Regression Analysis

In essence, we assume that the variance, skewness, and kurtosis do not vary from one subpopulation $(X = x_1)$ to the next $(X = x_2)$ under the assumed distribution. However, it is often the case that the variance of $Y|X$ at each subpopulation is not constant across all subpopulations. Instead, we often find larger variation than expected. The joint modeling of the mean and dispersion addresses such deviations from the model assumptions. In a like manner, one can also identify deviations in the skewness and kurtosis that may be present in the data.

4.2.3 Joint Modeling of the Mean and Dispersion

In a letter to the editor, Nelder et al. (1998) presented the joint modeling of the mean and dispersion by using two interlinked generalized linear models (GLM) for the mean and the dispersion. They use the deviance component, a function of Y and $\mu_{Y|X}$, as

the response for the dispersion model, and the inverse of the fitted values for the dispersion model as the prior weights for the mean model. An iterative-reiterative procedure is used to fit the mean and the dispersion submodels until both estimators of the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (in the mean and dispersion, respectively) converge. Nelder et al. (1998) used a GLM approach for fitting the dispersion submodel and as such used model-checking techniques for generalized linear models. Such techniques are applied directly to both submodels of the joint modeling of the mean and dispersion (McCullagh and Nelder 1989).

Consider a random variable Y_i for $i = 1, \dots, n$, that follows a distribution that is a member of the exponential family, so the log-likelihood function based on known parameters θ_i, ϕ_i, and w_i is $l(\theta_i, \phi_i^{-1}, w_i : y_i) = \sum_i [w_i \phi_i^{-1}\{y_i\theta_i - b(\theta_i)\} + c(y_i, w_i\phi_i^{-1})]$. For ϕ_i and w_i unknown but fixed, this form is a member of the exponential family if we assume that $\mathrm{c}(y_i, w_i\phi_i^{-1}) = -w_i\phi_i^{-1}a(y_i) - \frac{1}{2}s(-w_i\phi_i^{-1}) + t(y_i)$. Therefore, it follows that

$$l(\theta_i, \phi_i^{-1}, w_i : y_i) = \sum_i \left[w_i\phi_i^{-1}\{y_i\theta_i - b(\theta_i) - a(y_i)\} - \frac{1}{2}s(-w_i\phi_i^{-1}) + t(y_i)\right]. \quad (4.2)$$

Thus, we have a generalized linear model such that for $\mu_i = E(Y_i) = b'(\theta_i)$, $g_M(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$, where $\boldsymbol{x}_i' = (x_{i0}, \dots, x_{ip})$ is the vector of covariates, $\boldsymbol{\beta}$ is the vector of regression parameters, and $b'(\cdot)$ is the first derivative of the function $b(\theta_i)$ for link function $g_M(\cdot)$. The functions $a(y)$ and $b(\theta_i)$ are known functions. The variance is $var(Y_i) = \sigma_i^2 = \{w_i\phi_i^{-1}v(\mu_i)\}$, where $v(\mu_i) = b''(\theta_i)$ and $b''(\cdot)$ is the second derivative of $b(\cdot)$. Smyth (1989) reported that Barndorff-Nielsen and Blaesid (1983a; 1983b) studied generalized linear models and showed that $a(\cdot)$, $b(\cdot)$, and θ are related. Thus, we can think of the

likelihood function $l(\theta_i, \phi_i^{-1}, w_i: y_i)$ as a function of $\boldsymbol{\beta}$ through θ_i when ϕ_i^{-1} and w_i are held fixed and defines a submodel corresponding to the mean through $\boldsymbol{\beta}$.

Consider (4.2) and define a random variable $D_i = d_i$ for $i = 1, \dots, n$ for the dispersion d_i such that $d_i(y_i, \mu_i) = -2\{y_i\theta_i - b(\theta_i) - a(y_i)\}$. Substituting $d_i(y_i, \mu_i)$ and reparametrizing, we obtain a distribution of the form of the exponential family, so the log-likelihood function is of the form

$$l(\phi_i^{-1}, w_i: d_i) = \sum_i \left[\frac{1}{2}\{-w_i\phi_i^{-1}d_i - s(-w_i\phi_i^{-1})\} + t(y_i)\right].$$

This representation is similar to the form of the log-likelihood defined in the mean submodel and as such follows a generalized linear model with $E(D_i = d_i) = \delta_i = w_i\dot{s}(-w_i\phi_i^{-1})$ and $var(D_i) = 2w_i^2\ddot{s}(-w_i\phi_i^{-1})$. The dispersion submodel can be formulated with link function $g_V(\delta_i) = \boldsymbol{z_i'\gamma}$ where $\boldsymbol{z_i'} = (z_{i0}, \dots, z_{ip})$ is the vector of covariates and $\boldsymbol{\gamma}$ is the vector of regression parameters used to explain the dispersion. Thus, we can think of the likelihood function $l(\phi_i^{-1}, w_i: d_i)$ as a function of $\boldsymbol{\gamma}$ through ϕ_i^{-1} when w_i are held fixed as defining a submodel corresponding to the dispersion through $\boldsymbol{\gamma}$. It is common to model the variance as a function of the mean and allow the dispersion parameter to be dependent on certain covariates through unknown parameters.

4.3 Model Fit and Trimming Using Skewness and Kurtosis

4.3.1 Joint Modeling of Mean, Variance, Skewness, and Kurtosis

We define the standardized residual as $\eta = \epsilon/\sigma$ where $\epsilon = y - \mu$, similar to time series applications for incorporating skewness and kurtosis (Harvey and Siddique 1999; León, Rubio, and Serna 2005) in an extension to joint modeling. Then $E(\eta) = 0$, $E(\eta^2) = 1$,

$$E(\eta^3) = E\left[\left(\frac{\epsilon}{\sigma}\right)^3\right] = \frac{E\left[\left(Y-E(Y)\right)^3\right]}{\sigma^3} = \gamma_1,$$

and $E(\eta^4) = \gamma_2 + 3$. For the skewness submodel, let the response be

$$S = \eta^3 - \gamma_1$$

with mean $E(S) = \tau = 0$ and variance $Var(S) = Var(\eta^3) = E(\eta^6) - \gamma_1^2$. For the kurtosis submodel, let the response be

$$K = (\eta^4 - 3) - \gamma_2$$

with mean $E(K) = \kappa = 0$ and variance $Var(K) = Var(\eta^4) = E(\eta^8) - (\gamma_2^2 + 3)^2$. These additional submodels identify covariates which impact the deviations in the expected skewness and expected kurtosis under the assumed distribution.

The four submodels model the mean, dispersion, deviation from skewness, and deviation from kurtosis denoted as μ, ϕ, τ, and κ, respectively. We fit a mean submodel on the parameter μ and obtain the deviance to fit the dispersion submodel. We consider the standardized residuals from the mean submodel with the estimated variance to model the deviations in the skewness and kurtosis. Two-step generalized method of moments (GMM) is used to estimate the model coefficients and obtain standard errors for the skewness and kurtosis submodels. The estimation of the skewness and kurtosis submodel parameters do not require distributional assumptions. The moment conditions for each of the submodels are given in Sections 4.3.1.3 and 4.3.1.4.

4.3.1.1 Mean Submodel

Let $\boldsymbol{y} = (y_1, \dots, y_n)'$ be a random vector of observations from m subpopulations of X, and denote the link function $g_M(\cdot)$ such that for $E(Y|X) = \mu_{Y|X}$, we have the generalized linear model

$$g_M(\mu_i) = \boldsymbol{x_i}'\boldsymbol{\beta}$$

where the covariates $\boldsymbol{x_i'} = (x_{i0}, \dots, x_{ip})$ and the vector of regression coefficients $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)'$ with $var(y_i|\boldsymbol{X}) = \sigma_{ii}$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$. Thus, this is a mean submodel with covariates $\boldsymbol{X}$ and regression coefficients $\boldsymbol{\beta}$. The deviance measure is a function of $Y|X$ and $\mu_{Y|X}$,

$$D = \sum_{i=1}^{n} 2w_i\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\}$$

where $\theta(\cdot)$ is the canonical link, $\hat{\theta} = \theta(\mu)$, $\tilde{\theta} = \theta(y)$, and $b(\cdot)$ is the cumulant function obtained from the exponential family (McCullagh and Nelder 1989). The deviances $d_i(y_i, \mu_i)$ are used to model the variance.

4.3.1.2 Variance Submodel

Let the vector of responses be the deviance $\boldsymbol{d} = (d_1, \dots, d_n)'$ with mean $\boldsymbol{\phi}$ and link function $g_V(\cdot)$ such that the dispersion submodel is a generalized linear model,

$$g_V(\phi_i) = log(\phi_i) = \boldsymbol{z_i}'\boldsymbol{\gamma}$$

with covariates $\boldsymbol{z_i'} = (z_{i0}, \dots, z_{iq})$, regression parameters $\boldsymbol{\gamma} = (\gamma_0, \dots, \gamma_q)$, $var(d_i) = \sigma_{d_{ii}}$ and covariance matrix $\boldsymbol{\Omega} = \left(\sigma_{d_{ij}}\right)$. Fitting the variance submodel provides an estimate of the variance,

$$\widehat{\sigma^2} = \frac{\hat{\phi}}{w} V(\hat{\mu}_{Y|X}),$$

where w is the prior weights, $\hat{\phi}$ is the estimated dispersion parameter from the dispersion submodel, and $\hat{\mu}_{Y|X}$ is the estimated mean from the mean submodel. In Smyth (1989), the mean and variance submodels are fit using likelihood approaches. However, there are

other fitting techniques including restricted maximum likelihood (Smyth and Verbyla 1999).

4.3.1.3 Skewness Submodel

Recall we define the standardized residual $\eta = \epsilon/\sigma$ such that the estimated skewness is

$$g_1 = \hat{\eta}^3 = \frac{\hat{\epsilon}^3}{\hat{\sigma}^3} = \frac{\hat{\epsilon}^3}{(\hat{\sigma}^2)^{3/2}}.$$

We provide a generalized linear model to fit the skewness submodel through the deviation in skewness τ. Let $\boldsymbol{s} = (s_1, \dots, s_n)'$ be the deviations in skewness for $\boldsymbol{s} = \boldsymbol{g_1} - \hat{\boldsymbol{\gamma}}_{\boldsymbol{1}}$, where $\boldsymbol{g_1}$ is the estimated skewness defined above and $\boldsymbol{\gamma_1}$ is the expected skewness under the assumed distribution. Consider the link function $g_S(\cdot)$, which relates the vector of covariates $\boldsymbol{T}$ to the deviation in skewness. Thus, we have

$$g_S(\tau_i) = \tau_i = \boldsymbol{t_i}'\boldsymbol{\alpha}$$

where $\boldsymbol{t_i'} = (t_{i0}, \dots, t_{ir})$ and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_r)$ with $var(s_i) = \sigma_{s_{ii}}$ and covariance matrix $\boldsymbol{\Theta} = \left(\sigma_{s_{ij}}\right)$.

The regression parameters $\boldsymbol{\alpha}$ are estimated using a generalized method of moments approach. Let the GMM estimator $\hat{\boldsymbol{\alpha}}_{GMM}$ be the argument that minimizes the quadratic objective function $\boldsymbol{f}_n(\boldsymbol{\alpha})'\boldsymbol{W}_n\boldsymbol{f}_n(\boldsymbol{\alpha})$, such that

$$\hat{\boldsymbol{\alpha}}_{GMM} = argmin_{\alpha}\{\boldsymbol{f}_n(\boldsymbol{\alpha})'\boldsymbol{W}_n\boldsymbol{f}_n(\boldsymbol{\alpha})\}.$$

The sample moment conditions $\boldsymbol{f}_n(\boldsymbol{\alpha})$ are obtained using the empirical estimate of the population moment conditions

$$E\left[\frac{\partial\tau_i(\boldsymbol{\alpha})}{\partial\alpha_j}\{s_i - \tau_i(\boldsymbol{\alpha})\}\right] = 0.$$

In the case of an identity link, $\frac{\partial \tau_i(\boldsymbol{\alpha})}{\partial \alpha_j} = [1 \quad t_1 \quad \dots \quad t_r]'$ for r covariates. The weight matrix is estimated as

$$\widehat{\boldsymbol{W}}_n = \left[\frac{1}{n}\sum_{i=1}^{n} f(s_i, \widehat{\boldsymbol{\alpha}}) f(s_i, \widehat{\boldsymbol{\alpha}})'\right]^{-1}$$

using two-step generalized method of moments estimation with an identity weight matrix in the first step (Hansen 1982). The asymptotic variance of $\widehat{\boldsymbol{\alpha}}_{GMM}$ is computed as

$$\boldsymbol{V}_{GMM} = \frac{1}{n}[\boldsymbol{\Gamma}'\boldsymbol{W}\boldsymbol{\Gamma}]^{-1}$$

where $\boldsymbol{\Gamma}$ is the expected value of the Jacobian of population moment conditions, $E\left[\frac{\partial \boldsymbol{f}(s,\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right]$, evaluated at $\boldsymbol{\alpha} = \widehat{\boldsymbol{\alpha}}_{GMM}$ (Imbens and Spady 2002).

4.3.1.4 Kurtosis Submodel

Let

$$g_2 = \hat{\eta}^4 - 3 = \frac{\hat{\epsilon}^4}{\hat{\sigma}^4} - 3 = \frac{\hat{\epsilon}^4}{(\hat{\sigma}^2)^2} - 3.$$

be an estimate of the kurtosis. For the expected kurtosis under the assumed distribution γ_2, we evaluate the deviation in kurtosis κ through a generalized linear model. For deviations in the kurtosis $\boldsymbol{k} = (k_1, \dots, k_n)'$, where $\boldsymbol{k} = \boldsymbol{g_2} - \widehat{\boldsymbol{\gamma}}_{\boldsymbol{2}}$, and link function g_K,

$$g_K(\kappa_i) = \kappa_i = \boldsymbol{u_i'}\boldsymbol{\delta}$$

where $\boldsymbol{u_i}' = (u_1, \dots, u_c)$ are the covariates and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_c)$ are the regression parameters with $var(k_i) = \sigma_{k_{ii}}$ and covariance matrix $\boldsymbol{\Psi} = \left(\sigma_{k_{ij}}\right)$.

Similar to the skewness submodel, the GMM parameter estimates for the regression parameters are obtained as

$$\widehat{\boldsymbol{\delta}} = argmin_{\delta}\{\boldsymbol{f}_n(\boldsymbol{\delta})'\boldsymbol{W}_n\boldsymbol{f}_n(\boldsymbol{\delta})\}$$

using weight matrix

$$\widehat{\boldsymbol{W}}_n = \left[\frac{1}{n}\sum_{i=1}^{n} f\left(k_i, \widehat{\boldsymbol{\delta}}\right) f\left(k_i, \widehat{\boldsymbol{\delta}}\right)'\right]^{-1}$$

The population moment conditions to estimate $\widehat{\boldsymbol{\delta}}_{GMM}$ in the kurtosis submodel are

$$E\left[\frac{\partial \kappa_i(\boldsymbol{\delta})}{\partial \delta_j}\{k_i - \kappa_i(\boldsymbol{\delta})\}\right] = 0$$

with the asymptotic variance for $\boldsymbol{\delta}$ calculated as described above. Thus, the vectors of regression parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\delta}$ are estimated using the four submodels (mean, variance, skewness, and kurtosis). The link function for the mean model is selected based on the scale of the responses. A log link is used for the dispersion submodel. Nelder et al. (1998) reported that sufficient data is needed to discriminate between alternative dispersion links. We use the identity link function for the skewness and kurtosis submodels, as the identity link function performs well for both models and is appropriate for negative values. The four submodels are summarized in Table 4.2.

Table 4.2. Four Interlinked Submodels (Mean, Variance, Skewness, and Kurtosis)

Component	**Mean**	**Dispersion**	**Deviation in Skewness**	**Deviation in Kurtosis**
Response	Y_i	D_i	S_i	K_i
Mean	μ	ϕ	τ	κ
Link	$g_M(\cdot)$	$g_V(\cdot)$	$g_S(\cdot)$	$g_K(\cdot)$
Predictor	X_i	Z_i	T_i	U_i
Parameter	β	γ	α	δ
Covariance	Σ	Ω	Θ	Ψ

4.3.2 Trimming Using Skewness and Kurtosis

The presence of outliers impacts the model fit as well as the parameter estimates. Extreme values, due to the presence of skewness and kurtosis, can also impact the model fit. Neykov, et al. (2012) showed that trimming in joint modeling can identify outliers and reduce bias in the parameter estimates. As such, we recommend the removal of outliers based on the skewness and kurtosis model outcomes, S and K. Suspected outliers can be removed using the cutoff $\left|\frac{S}{K}\right| > 2$. This utilizes the ratio of the deviation in skewness and deviation in kurtosis where

$$\frac{S}{K} = \frac{\eta^3 - \gamma_1}{(\eta^4 - 3) - \gamma_2}.$$

This ratio identifies outliers due to the skewness or kurtosis. As we show in the simulation study, the removal of values with $\left|\frac{S}{K}\right|$ greater than 2 results in an improved mean-dispersion model fit and the model parameter estimates are similar or closer to the true values.

4.4 Simulation Study

A simulation study is conducted to demonstrate the benefits of fitting additional submodels. The study consists of normal and gamma distributed data generated under certain conditions, each replicated 1000 times. We simulate each data set with 500 observations. The joint modeling is used to evaluate the impact of the covariates on the mean, variance, skewness, and kurtosis.

4.4.1 Normal Data

We evaluate normally distributed data under four conditions. Under the normal distribution, y has a mean of μ, variance of $\phi\sigma^2$, skewness of 0 and kurtosis of 0. For each simulation condition below, the mean is generated as $\mu_i = 1 + x_i$ where $x_i \sim N(0,1)$.

Condition 1: Let ϕ = 1 (homoscedastic normal).

Condition 2: Let $log(\phi_i) = -2 - 2x_i$ for $i = 1, \ldots, 500$ (heteroscedastic normal).

Condition 3: Let the data be simulated from the R package *sn* (Azzalini 2017) with parameters $\xi = 1 + x_i$, $\omega = \sqrt{\phi}$, $\alpha = 1$, and $\tau = 0$ where $\log(\phi_i) = -2 - 2x_i$ (heteroscedastic skew normal).

Condition 4: Let the data be simulated from the uniform distribution, $y_i \sim Uniform(a, b)$ with $a = -\sqrt{3} + \mu$ and $b = 2\mu - a$, which has an expected kurtosis of -1.2 (Chissom 1970).

The average parameter estimates and standard errors for each of the four conditions are provided in Table 4.3. The percentage of simulations in which the result is statistically significant for α = 0.05 is also reported.

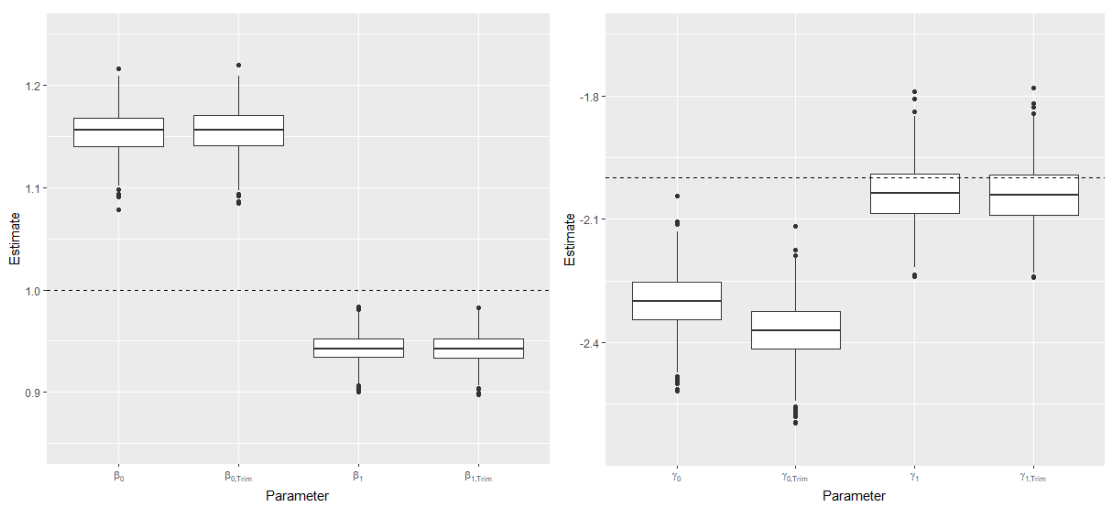
Table 4.3. Simulation Results for Normal Data

	Mean		Dispersion		Deviation in Skewness		Deviation in Kurtosis	
	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\gamma_0}$	$\boldsymbol{\gamma_1}$	$\boldsymbol{\alpha_0}$	$\boldsymbol{\alpha_1}$	$\boldsymbol{\delta_0}$	$\boldsymbol{\delta_1}$
Condition 1								
Estimate	0.9993	0.9993	-0.0098	-0.0020	-0.0025	-0.0018	-0.0285	-0.0008
Std. Error	0.0447	0.0447	0.0953	0.0955	0.1686	0.1643	0.4120	0.3889
% Significant	100.0%	100.0%	0.5%	0.4%	0.1%	0.0%	0.1%	0.0%
Condition 2								
Estimate	1.0001	0.9998	-2.0099	-2.0087	-0.0034	-0.0001	-0.0266	0.0019
Std. Error	0.0149	0.0074	0.0953	0.0955	0.1696	0.1633	0.4125	0.3879
% Significant	100.0%	100.0%	100.0%	100.0%	0.3%	2.6%	0.0%	0.0%
Condition 3								
Estimate	1.1543	0.9426	-2.2993	-2.0374	0.5154	-0.5737	0.0807	-0.0329
Std. Error	0.0126	0.0062	0.0963	0.0965	0.1753	0.1982	0.4578	0.5071
% Significant	100.0%	100.0%	100.0%	100.0%	91.7%	90.4%	0.0%	0.0%
Condition 4								
Estimate	0.0040	-0.0007	0.0093	0.0001	0.0130	0.0010	-1.2161	0.0014
Std. Error	0.0451	0.0452	0.0725	0.0727	0.0872	0.0873	0.1068	0.1084
% Significant	4.2%	5.0%	0.1%	0.1%	0.9%	0.3%	100.0%	0.0%

For the first three conditions, we expect the mean regression parameter estimates to be close to 1. Under Condition 1, the dispersion parameters should be close to 0 while in Condition 2, the dispersion parameters should both be equal to -2. The parameters for the skewness and kurtosis submodels are expected to be equal to or very close to zero. We find that these conditions hold and deviations in the skewness and kurtosis are not identified. Under Condition 3, which is a heteroscedastic skewed normal distribution, we expect to identify significant deviation in skewness and see that the regression parameters (α_0, α_1) for the skewness submodel are significant in 91.7% and 90.4% of cases, respectively. Under Condition 4, the kurtosis submodel identifies a significant constant difference in the kurtosis. The average estimate of the intercept parameter for the deviation in kurtosis submodel is $\hat{\delta}_0$ = -1.2161 and is significant in 100.0% of the 1000

simulations. This represents the difference between the expected kurtosis of the normal distribution and the expected kurtosis of the uniform distribution.

As there are highly significant deviations in skewness under Condition 3, we trim the data based on the cut point for the ratio $\left|\frac{S}{K}\right| > 2$. Trimming improved the model fit for 99.9% of the simulated data sets under Condition 3. The mean and variance parameter estimates for the original and trimmed models are shown in Figure 4.3. The true simulated value is denoted by the dashed line. In this case, the parameter estimates from the original and trimmed models are comparable.



(a) (b)

Figure 4.3. Original and Trimmed Estimates of the Normal a) Mean and b) Variance Model Parameters

4.4.2 Gamma Data

We simulate data under three conditions using a gamma distribution with scale parameter $\phi_i\mu_i$ and shape parameter ϕ_i^{-1}. Thus y_i has mean μ_i, variance $\phi_i\mu_i^2$, skewness

$2\phi_i^{1/2}$, and kurtosis $6\phi_i$. In each condition, the covariate x_i is uniformly distributed on (0, 1) and the mean is generated based on $log(\mu_i) = 1 + x_i$ such that $\beta_0 = \beta_1 = 1$.

Condition 1: Let ϕ = 1.

Condition 2: Let $log(\phi_{i|X}) = -2 - 2x_i$ for $i = 1, \ldots, 500$.

Condition 3: Let $log(\phi_{i|X}) = -2 - 2x_i$ for $i = 1, \ldots, 500$ and ten percent of the data be randomly selected and perturbed as (0.25 + 0.5x_i)%.

The estimated parameter estimates, standard errors, and percent of statistically significant results are shown in Table 4.4.

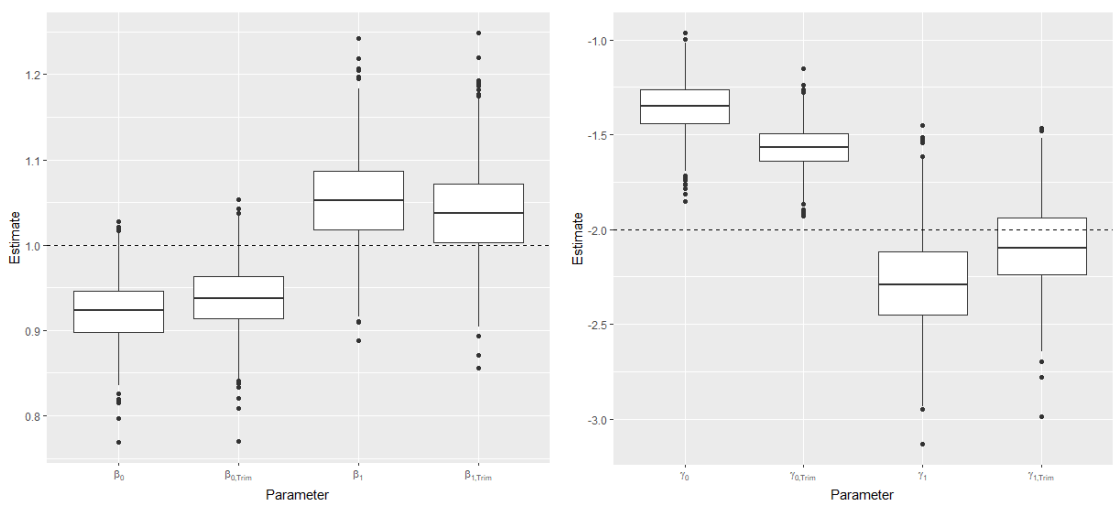
Table 4.4. Simulation Results for Gamma Data

	Mean		**Dispersion**		**Deviation in Skewness**		**Deviation in Kurtosis**	
	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\gamma_0}$	$\boldsymbol{\gamma_1}$	$\boldsymbol{\alpha_0}$	$\boldsymbol{\alpha_1}$	$\boldsymbol{\delta_0}$	$\boldsymbol{\delta_1}$
Condition 1								
Estimate	0.9909	1.0095	-0.0035	-0.0065	0.0883	-0.1968	0.3856	-1.1858
Std. Error	0.1211	0.1711	0.1933	0.3348	1.1771	2.0155	6.6177	11.2504
% Significant	100.0%	99.9%	0.2%	0.1%	7.1%	0.6%	15.5%	0.9%
Condition 2								
Estimate	0.9998	1.0002	-2.0044	-2.0060	-0.0050	-0.0113	-0.0249	-0.0458
Std. Error	0.0306	0.0385	0.1892	0.3293	0.4279	0.6980	1.3494	2.1144
% Significant	100.0%	100.0%	100.0%	100.0%	5.9%	1.5%	10.0%	0.9%
Condition 3								
Estimate	0.9224	1.0524	-1.3522	-2.2844	-0.7327	0.3757	-2.1286	1.9654
Std. Error	0.0388	0.0480	0.2066	0.3606	0.2814	0.5066	0.6886	1.2224
% Significant	100.0%	100.0%	100.0%	100.0%	73.0%	10.4%	80.5%	51.4%

Under Conditions 1 and 2, the mean estimates for the mean and dispersion model parameters are close to the true values ($\beta_0 = 1, \beta_1 = 1, \gamma_0 = 0, \gamma_1 = 0$ for Condition 1 and $\beta_0 = 1, \beta_1 = 1, \gamma_0 = -2, \gamma_1 = -2$ for Condition 2) while the parameter estimates for the skewness and kurtosis submodels are close to 0 and are not significant in most simulations. Under Condition 3, the perturbation adds additional skewness and kurtosis,

and so the skewness and kurtosis submodels have non-zero parameter estimates. The intercept parameter estimates (α_0 and δ_0) are found to be significant in more than 73% of cases, which indicates some differences in the skewness and the kurtosis as compared to the expected skewness and kurtosis under the gamma distribution.

Under Condition 3, the simulated data have significant skewness and kurtosis. However, trimming based on $\left|\frac{S}{K}\right| > 2$ resulted in an improvement in 100% of the mean and dispersion model fits. The parameter estimates in the mean and variance models after trimming the data were closer to the true parameter values as compared to the estimates prior to trimming (Figure 4.4).



(a) (b)

Figure 4.4. Original and Trimmed Estimates of the Gamma a) Mean and b) Variance Model Parameters

4.5 Numerical Example

While the accurate detection of breast cancer remains a challenge, recent studies have indicated that breast tumors are associated with fluctuations in serum protein biomarkers (SPB) and tumor associated antibodies (TAAb). This discovery has resulted in the consideration of proteomic assays as an additional diagnostic to breast cancer imaging. Henderson et al. (2016) integrated data from biomarker measurements in patients and found clinical sensitivity and specificity rates of breast cancer detection around 80% which support the use of proteomic approaches. We employ joint modeling of the mean, variance, skewness, and kurtosis to determine if any of the biomarker expression data are associated with more extreme probabilities of a breast cancer diagnosis.

We examine breast cancer screening data from 190 women (Henderson, et al. 2016) evaluated at Mercy Women's Center – Oklahoma City. Our outcome of interest is breast cancer diagnosis. Information collected includes demographic information as well as 57 SBP and TAAb measurements. We focus on the biomarkers basic fibroblast growth factor (bFGF), Fas Ligand (FasL), Interleukin 10 (IL-10), Interleukin 12 (IL-12), Interleukin 8 (IL-8), Placental Growth Factor (PIGF), Vascular endothelial growth factor subtype D (VEGF-D), and Cancer antigen 15.3 (CA15.3). We fit a logistic regression model which assumes a Bernoulli distribution with skewness $\frac{1-2p}{\sqrt{p(1-p)}}$ and kurtosis $\frac{6p^2-6p+1}{p(1-p)}$. The parameter estimates and standard errors for the models are displayed in Table 4.5. Statistically significant variables are shown in bold.

Table 4.5. Mean, Variance, Skewness, and Kurtosis Parameter Estimates (Standard Error) for Breast Cancer Prediction

Covariate	**Mean**	**Dispersion**	**Deviation in Skewness**	**Deviation in Kurtosis**
Intercept	**-3.049 (1.319)**	**1.714 (0.639)**	**-17.856 (0.027)**	**190555.108 (82.837)**
Age	**0.097 (0.021)**	1.81E-3 (9.17E-3)	-0.289 (0.827)	3029.108 (2591.318)
bFGF	**0.044 (0.021)**	**-0.035 (0.012)**	0.479 (0.654)	-667.660 (1961.532)
FasL	**-0.215 (0.051)**	**-0.0532 (0.023)**	-0.618 (1.885)	-5063.869 (5631.462)
IL-10	**-2.758 (0.444)**	**-1.884 (0.241)**	**136.950 (0.024)**	**19923.309 (72.334)**
IL-12	**-0.005 (0.002)**	8.41E-4 (1.11E-3)	-0.035 (0.104)	244.657 (362.027)
IL-8	**0.006 (0.001)**	3.68E-4 (7.55E-4)	**0.145 (0.026)**	-168.729 (87.909)
PIGF	-0.018 (0.027)	4.34E-3 (0.013)	-1.764 (1.150)	-318.056 (3329.274)
VEGF-D	**-0.001 (0.000)**	2.69E-5 (1.99E-4)	-0.015 (0.019)	86.088 (61.496)
CA15.3	**0.002 (0.000)**	**-1.02E-3 (1.17E-4)**	**0.203 (0.092)**	**-611.735 (307.860)**

The covariate age and most biomarkers are significant predictors of breast cancer diagnosis. In addition, the biomarkers bFGF and FasL significantly impact the dispersion. Two biomarkers, IL-10 and CA15.3, are drivers of the dispersion, skewness, and kurtosis. The positive coefficients in the skewness submodel indicate that larger measurements of these biomarkers are associated with extremely high probabilities of breast cancer diagnoses. The IL-10 biomarker has a positive kurtosis coefficient suggesting relatively low probabilities in the tails. The negative kurtosis coefficient for CA15.3 indicates that this biomarker is indicative of extreme probabilities, such as a very low probability or a very high probability of breast cancer. The model also suggests that IL-8 has a

significantly higher skewness than expected under the binomial distribution, indicating that higher values of IL-8 are associated with extremely high probabilities of a breast cancer diagnosis.

We trim the data based on the skewness and kurtosis ratio and find an improvement in the mean-dispersion model fit (-2 log likelihood of 150.623 vs 176.1547). Modeling the trimmed data indicates that the biomarker IL-8 is not as useful for predicting breast cancer as previously observed. In addition, FasL is not significant in the dispersion submodel. However, IL-8 has a significant impact on the variance of breast cancer diagnoses.

Table 4.6. Mean and Variance Parameter Estimates (Standard Error) for Breast Cancer Prediction after Trimming

Covariate	**Mean**	**Dispersion**
Intercept	**-4.501 (1.512)**	**2.116 (0.910)**
Age	**0.140 (0.024)**	-0.006 (0.013)
bFGF	**0.070 (0.022)**	**-0.061 (0.017)**
FasL	**-0.238 (0.054)**	-0.047 (0.031)
IL-10	**-3.083 (0.407)**	**-2.147 (0.327)**
IL-12	**-0.006 (0.003)**	9.42E-4 (1.51E-3)
IL-8	8.09E-5 (0.002)	**0.002 (0.001)**
PIGF	-0.056 (0.029)	0.014 (0.018)
VEGF-D	**-0.001 (0.000)**	-4.18E-5 (2.74E-4)
CA15.3	**0.002 (0.000)**	**-1.12E-3 (1.59E-4)**

4.6 Conclusions

In this paper, we expand on joint modeling to include submodels for the skewness and kurtosis. Inference on the skewness identifies significant deviations due to extreme observations in the tails. Inference on the kurtosis aids in the understanding of the concentration of the observations near the mean, to determine if the variability is due to a few extreme differences from the mean rather than a few modest differences from the mean. Simulation studies show that it is possible to simultaneously model the mean, variance, skewness, and kurtosis. In addition, we find that trimming based on the deviation in skewness and deviation in kurtosis leads to improved model fit and reliable

parameter estimates for the mean and dispersion submodels. We revisit breast cancer data for 190 women reported by Henderson, et al. (2016) and model breast cancer diagnosis. We identify covariates contributing to skewness or kurtosis, and can use this information to improve model fitting and to address outliers.

CHAPTER 5

CONCLUSIONS

Underlying distributional assumptions impact model estimation and fit. In particular, assumptions about the variance and higher order moments can lead to inaccurate parameter estimates and standard errors when the true variance in the data is not reflected in the approach. Methods to identify and appropriately account for deviations in the underlying distribution improve model fit and are particularly useful for correlated data.

The first paper presents a generalized method of moments approach through a canonical parameterization of the mean-variance relationship. The canonical parameterization generalizes the form to any distribution in the exponential family. In addition, the generalized method of moments approach allows both mean-variance parameters to be estimated simultaneously and does not require an underlying distributional assumption. The generalized method of moments estimation approach is shown to be computationally tractable and flexible. The simulation study confirms the estimation accuracy of the parameter estimates, which have small standard errors. A test is developed to determine if there is significant deviation in the mean-variance relationship.

The second paper implements the mean-variance relation parameterization in a modeling application. The mean-variance relationship is extended for estimation in two-level hierarchical data and is implemented in the covariance matrix for generalized quasi-likelihood modeling. This adjusted generalized quasi-likelihood approach accounts for the variance in the data and is reflected in the estimation of the random effect. An

evaluation of the approach through two examples demonstrates that in the case of correlated data, the parameter estimate for the random effect is improved by implementing this model as compared to a standard generalized quasi-likelihood model.

Additional deviations in moment assumptions are explored and addressed in the third paper. Joint modeling implements two simultaneous generalized linear models to improve model fit by modeling the unexplained variation in the mean model. We expand on this method and incorporate submodels to account for the skewness and kurtosis. This approach, the joint modeling of the mean, variance, skewness, and kurtosis, identifies and estimates the association of covariates with skewness and kurtosis in a particular set of data. Moreover, a cutoff based on the deviations in the skewness and kurtosis can remove outliers and improve the model fit of the mean and dispersion submodels. Parameter estimates in the trimmed model showed an improvement and had more accurate estimates.

Overall, this research addresses the underlying assumption about moments of the distribution. Three methods are proposed and tested to estimate and model the variance and higher order moments. Accounting for deviations in the estimated variance through models such as adjusted generalized quasi-likelihood or joint modeling of the mean, variance, skewness, and kurtosis improve the model accuracy and extends the understanding of the covariate associations with the shape of the distribution.

REFERENCES

Azzalini, Adelchi. 2017. *Package 'Sn'*. https://cran.r-project.org/web/packages/sn/sn.pdf.

Bai, Jushan, and Serena Ng. 2005. "Tests for Skewness, Kurtosis, and Normality for Time Series Data." *Journal of Business and Economic Statistics* 23, no. 1: 49-60.

Balanda, Kevin P., and H.L. MacGillivray. 1988. "Kurtosis: A Critical Review." *The American Statistician* 42, no. 2: 111-119.

Barndorff-Nielsen, Ole, and P Blaesild. 1983a. "Exponential Models with Affine Dual Foliations." *The Annals of Statistics* 11, no. 3: 753-769.

---. 1983b. "Reproductive Exponential Families." *The Annals of Statistics* 11, no. 3: 770-782.

Bera, Anil K., and Carlos M. Jarque. 1981. "Efficient Tests for Normality, Homoscedasticity and Serial Indepdence of Regression Residuals: Monte Carlo Evidence." *Economics Letters* 7, no. 4: 313-318.

Bhargava, Alok. 1994. "Modelling the Health of Filipino Children." *Journal of the Royal Statistical Society Series A* 157, no. 3: 417-432.

Breslow, N.E., and D.G. Clayton. 1993. "Approximate Inference in Generalized Linear Mixed Models." *Journal of the American Statistical Association* 88, no. 421: 9-25.

Chissom, Brad S. 1970. "Interpretation of the Kurtosis Statistic." *The American Statistician* 24, no. 4: 19-22.

Cox, David Roxbee. 1983. "Some Remarks on Overdispersion." *Biometrika* 70, no. 1: 269-274.

Cox, David Roxbee, and Nancy Reid. 1987. "Parameter Orthogonality and Approximate Conditional Inference." *Journal of the Royal Statistical Society Series B* 49, no. 1: 1-39.

Dean, CB. 1992. "Testing for Overdispersion in Poisson and Binomial Regression Models." *Journal of the American Statistical Association* 87: 451-457.

DeCarlo, Lawrence T. 1997. "On the Meaning and Use of Kurtosis." *Psychological Methods* 2, no. 3: 292-307.

Dobson, Annette J, and Adrian G Barnett. 2008. *An Introduction to Generalized Linear Models, Third Edition*. New York: CRC Press.

Donald, Stephen G, Guido W Imbens, and Whitney K Newey. 2009. "Choosing Instrumental Variables in Conditional Moment Restriction Models." *Journal of Econometrics* 152: 28-36.

Fogler, H. Russell, and Robert C. Radcliffe. 1974. "A Note on Measurement of Skewness." *The Journal of Financial and Quantitative Analysis* 9, no. 3: 485-489.

Groeneveld, Richard A., and Glen Meeden. 1984. "Measuring Skewness and Kurtosis." *Journal of the Royal Statistical Society Series D* 33, no. 4: 391-399.

Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50, no. 4: 1029–1054.

Harris, Kathleen Mullan, et al. 2009. *The National Longitudinal Study of Adolescent to Adult Health: Research Design*. http://www.cpc.unc.edu/projects/addhealth/design.

Harvey, Cambell R., and Akhtar Siddique. 1999. "Autoregressive Conditional Skewness." *The Journal of Financial and Quantitative Analysis* 34, no. 4: 465-487.

Henderson, Meredith C., et al. 2016. "Integration of Serum Protein Biomarker and Tumor Associated Autoantibody Expression Data Increases the Ability of a Blood-Based Proteomic Assay to Identify Breast Cancer." *PLOS One* 11, no. 8. http://dx.doi.org/10.1371/journal.pone.0157692.

Imbens, Guido W, and Richard Spady. 2002. "Confidence Intervals in Generalized Method of Moments Models." *Journal of Econometrics* 107, no. 87-98.

Jarque, Carlos M., and Anil K. Bera. 1980. "Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals." *Economics Letters* 6, no. 3: 255-259.

Jiang, Jiming. 2003. "Empirical Method of Moments and Its Applications." *Journal of Statistical Planning and Inference* 115, no. 1: 69-84.

Joanes, D.N., and C.A. Gill. 1998. "Comparing Measures of Sample Skewness and Kurtosis." *Journal of the Royal Statistical Society Series D* 47, no. 1: 183-189.

Kendall, Maurice G., and Alan Stuart. 1969. *The Advanced Theory of Statistics*. New York: McGraw Hill.

Kukush, Alexander, Andrii Malenko, and Hans Schneeweiss. 2009. "Optimality of the Quasi-Score Estimator in a Mean-Variance Model with Applications to

Measurement Error Models." *Journal of Statistical Planning and Inference* 139: 3461-3472.

Lalonde, Trent L, Jeffrey R Wilson, and Jianqiong Yin. 2014. "Gmm Logistic Regression Models for Longitudinal Data with Time-Dependent Covariates and Extended Classifications." *Statistics in Medicine* 33, no. 27: 4756-4769.

Lee, Youngjo, and John A Nelder. 2000. "Two Ways of Modelling Overdispersion in Non-Normal Data." *Journal of the Royal Statistical Society Series C* 49, no. 4: 591-598.

---. 2006. "Double Hierarchical Generalized Linear Models." *Journal of the Royal Statistical Society Series C* 55, no. 2: 139-185.

León, Ángel, Gonzalo Rubio, and Gregorio Serna. 2005. "Autoregresive Conditional Volatility, Skewness and Kurtosis." *The Quarterly Review of Economics and Finance* 45: 599-618.

Liang, Kung-Yee, and Scott L Zeger. 1986. "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 73, no. 1: 13-22.

McCullagh, Peter, and John A Nelder. 1989. *Generalized Linear Models, 2nd Ed*. London: Chapman and Hall.

McCullagh, Peter, and Robert Tibshirani. 1990. "A Simple Method for the Adjustment of Profile Likelihoods." *Journal of the Royal Statistical Society Series B* 52, no. 2: 325-344.

Milanzi, Elasma, Ariel Alonso, and Geert Molenberghs. 2012. "Ignoring Overdispersion in Hierarchical Loglinear Models: Possible Problems and Solutions." *Statistics in Medicine* 31, no. 14 (June 30, 2012): 1475-1482. http://dx.doi.org/10.1002/sim.4482.

Morel, Jorge G, and Nagaraj K Neerchal. 2012. *Overdispersion Models in Sas*. Cary: SAS Institute Inc.

Nelder, John A, et al. 1998. "Joint Modeling of Mean and Dispersion." *Technometrics* 40, no. 2: 168-175.

Neykov, N.M., P. Filzmoser, and P.N. Neytchev. 2012. "Robust Joint Modeling of Mean and Dispersion through Trimming." *Computational Statistics & Data Analysis* 56, no. 1: 34-48.

Pack, Simon E. 1986. "Hypothesis Testing for Proportions with Overdispersion." *Biometrics* 42, no. 4: 967-972.

Rencher, Alvin C., and G. Bruce Schaalje. 2008. *Linear Models in Statistics*. Hoboken: John Wiley & Sons.

Royall, Richard, and Tsung-Shan Tsou. 2003. "Interpreting Statistical Evidence by Using Imperfect Models: Robust Adjusted Likelihood Functions." *Journal of the Royal Statistical Society Series B* 65, no. 2: 391-404.

Smyth, Gordon K. 1989. "Generalized Linear Models with Varying Dispersion." *Journal of the Royal Statistical Society Series B* 51, no. 1: 47-60.

Smyth, Gordon K, and Arunas P Verbyla. 1999. "Adjusted Likelihood Methods for Modelling Dispersion in Generalized Linear Models." *Environmetrics* 10, no. 6: 695-709.

Soberón, Alexandra, and Winfried Stute. 2017. "Assessing Skewness, Kurtosis and Normality in Linear Mixed Models." *Journal of Multivariate Analysis* 161: 123-140.

Sutradhar, Brajendra C. 2004. "On Exact Quasilikelihood Inference in Generalized Linear Mixed Models." *Sankhy-a: The Indian Journal of Statistics* 66, no. 2: 263-291.

Tan, Fei, et al. 2010. "A Full Likelihood Procedure of Exchangeable Negative Binomials for Modelling Correlated and Overdispersed Count Data." *Journal of Statistical Planning and Inference* 140: 2849-2859.

Tsou, Tsung-Shan. 2011. "Determining the Mean-Variance Relationship in Generalized Linear Models—a Parametric Robust Way." *Journal of Statistical Planning and Inference* 141: 197-203.

Wang, Bei, and Jeffrey R Wilson. 2017. "Comparative GMM and GQL Logistic Regression Models on Hierarchical Data." *Journal of Applied Statistics* 45, no. 3: 409-425.

Wedderburn, RWM. 1974. "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method." *Biometrika* 61, no. 3: 439-447.

Weisberg, Sanford. 1985. *Applied Linear Regression*. 2nd ed., Wiley Series in Probability and Mathematical Statistics Applied Probability and Statistics ,. New York: Wiley.

Wilson, Jeffrey R, and Kenneth J Koehler. 1991. "Hierarchical Models for Cross-Classified Overdispersed Multinomial Data." *Journal of Business and Economic Statistics* 9, no. 1: 103-110.

Wilson, Jeffrey R, and Kent A Lorenz. 2015. *Modeling Binary Correlated Responses Using Sas, Spss and R*. New York: Springer.

Xiang, Liming, et al. 2007. "A Score Test for Overdispersion in Zero-Inflated Poisson Mixed Regression Model." *Statistics in Medicine* 26: 1608-1622.

Yang, Zhao, James W Hardin, and Cheryl L Addy. 2009. "A Note on Dean's Overdispersion Test." *Journal of Statistical Planning and Inference* 139: 3675-3678.

Zsohar, Peter. 2012. "Short Introduction to the Generalized Method of Moments." *Hungarian Statistical Review Special Number 16*.