DIGITALCOMMONS
@WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 11 | Issue 1

Article 19

5-1-2012

Regression Models for Mixed Over-Dispersed Poisson and Continuous Clustered Data: Modeling BMI and Number of Cigarettes Smoked Per Day

Folefac Atem
Brigham and Women's Hospital, Boston, MA

Julius S. Ngwa
Boston University

Abidemi Adeniji
University of Pittsburgh

Recommended Citation

Atem, Folefac; Ngwa, Julius S.; and Adeniji, Abidemi (2012) "Regression Models for Mixed Over-Dispersed Poisson and Continuous Clustered Data: Modeling BMI and Number of Cigarettes Smoked Per Day," *Journal of Modern Applied Statistical Methods*: Vol. 11: Iss. 1, Article 19.
Available at: http://digitalcommons.wayne.edu/jmasm/vol11/iss1/19

Journal of Modern Applied Statistical Methods
May 2012, Vol. 11, No. 1, 218-232

1538 – 9472/12/$95.00

Regression Models for Mixed Over-Dispersed Poisson and Continuous Clustered Data: Modeling BMI and Number of Cigarettes Smoked Per Day

Folefac Atem
Brigham and Women's Hospital,
Boston, MA

Julius S. Ngwa
Boston University,
Boston, MA

Abidemi Adeniji
University of Pittsburgh,
Pittsburgh, PA

Clustered data, multiple observations collected on the same experimental unit, is common in epidemiological studies. Bivariate outcome data is often the result of interest in two correlated response variables. An efficient method is presented for dealing with bivariate outcomes when one outcome is continuous and the other is a count using a simple transformation to handle over-dispersed Poisson data. A multilevel analysis was performed on data from the National Health Interview Survey (NHIS) with body mass index (BMI) and the number of cigarettes smoked per day (NCS) as responses. Results show that these random effects models yield misleading results in cases where the data is not transformed.

Key words: Joint bivariate model, random effect model, multilevel, mixed model, GLIMMIX.

Introduction

Modeling bivariate outcome using joint multivariate random effect models (JMRE) is a popular approach in the medical field. There are a number of conditions where a disease under study is well understood when two outcomes are considered. For example, in clinical trials the clinician may be interested in the joint evolution of HIV RNA and CD4+t lymphocytes in a cohort of HIV-1 infected patients treated with active antiretroviral drugs. Bellamy (1995) studied the study of the risk factors associated with the progression of osteoarthritis (OA) of the knee. In this study two outcomes were collected, the Western Ontario and McMaster Universities (WOMAC) disability score and the number of missed work days for the past three months due to knee pain. The JMRE model allows the modeling of mixed effects and bivariate outcomes.

There are several advantages of modeling bivariate outcomes (see Laird and Ware, 1982; Dempster, et al., 1984; Bagiella, 2000; Pantazis & Touloumi, 2007; McCulloch, 2008; Atem, et al,. 2010). First, this model allows exploration of variations at different levels of a hierarchy and modeling of a correlation structure serially and across two outcomes. Second, the bivariate JMRE model achieves greater bias reduction in all model parameters compared to the two independent JMRE models. Third, there is greater flexibility in dealing with exploratory variables; the JMRE model can conveniently test hypotheses on either end point individually or simultaneously. JMRE can also handle missing data as the parameter estimates are obtained by techniques of maximum likelihood; the variance and means of the outcomes are estimated, thus the choice of the variance must be taken into consideration.

Rao (1973) suggested that irrespective of the chosen variance, the fixed effect estimates are unbiased; however, the estimates can achieve maximum efficiency only when the appropriate variance is specified. Lastly, multiple testing can be avoided by forming joint models without resorting to ad hoc methods such as the Bonferroni adjustment or using advanced methods such as those as presented by Dmitrienko, et al. (2009). Moreover, because the bivariate model is a regression model,

Folefac Atem is a Biometric Consultant. Email him at: folefac_atem@yahoo.com. Julius S. Ngwa is a Ph.D. candidate in the Biostatistics Department at the School of Public Health. Abidemi K. Adeniji has a Ph.D. in biostatistics and is a researcher at The Epidemiology Data Center (EDC), University of Pittsburgh.

classification and continuous predictors can be incorporated; the predictors can either be time variant or invariant.

This study develops a straightforward and efficient method to handle bivariate outcome data where one outcome is continuous and the other is count with over-dispersion; a test is proposed to identify possible over-dispersion in the count outcome. For modeling, data from the National Health Interview Survey (NHIS) from 1997-2006 was used to establish the joint relationship between body mass index (BMI; continuous) and number of cigarettes smoked per day (NCS; count), adjusting for race, gender and age. The National Health Interview Survey (NHIS) was founded in 1957 as an annual nationwide survey of approximately 40,000 households in civilian, non-institutionalized populations. It is conducted by the National Center for Health Statistics (NCHS) and administered by the United States Census Bureau.

Regression Techniques

Regression techniques for hierarchical data have been referred to in the literature under different names, including random coefficient models (Rao, 1965) and hierarchical linear model (Bryk & Raudenbush, 1987). Song, et al. (2008) refers to bivariate and multivariate analyses as structural level equations.

Bivariate Model

For the case of two continuous markers with correlated random effect superscripts a and b are used to distinguish the two markers. Subscripts j and i are used to denote the information of the j^{th} measurement for the i^{th} individual. Assuming that the marker trends can be explained by two linear mixed models with correlated random-effects, a bivariate model for the multilevel marker measurements can be presented as follows (Thiebaut, et al., 2002):

$$y_{ij}^a = x_{ij}^a \beta^a + z_{ij}^a \lambda_i^a + e_{ij}^a$$
$$\text{and} \qquad (1)$$
$$y_{ij}^b = x_{ij}^b \beta^b + z_{ij}^b \lambda_i^b + e_{ij}^b$$

Vectors $\lambda_i^a(k^a \times 1)$ and $\lambda_i^b(k^b \times 1)$ contain random (subject-specific) regression coefficients for k^a and k^b predictor variables included in the corresponding design vectors $Z_{ij}^a(1 \times k)$ and $Z_{ij}^b(1 \times k)$ respectively. The joint distribution of λ_i^a and λ_i^b is assumed to be multivariate normal with zero means and variance–covariance matrix:

$$\Sigma = \begin{pmatrix} \Sigma^a & \Sigma^{ab} \\ \Sigma^{ab} & \Sigma^b \end{pmatrix} \qquad (2)$$

Matrices $\Sigma^a(k^a \times k^a)$ and $\Sigma^b(k^b \times k^b)$ are variance and covariance of the random effects for both outcomes respectively, while $\Sigma^{ab}(k^b \times k^a)$ specifies the covariance structure of the random effects between outcomes.

Vectors $\beta^a(p^a \times 1)$ and $\beta^b(p^b \times 1)$ contain fixed regression coefficients for both markers respectively; $x_{ij}^a(1 \times p^a)$ and $x_{ij}^b(1 \times p^b)$ are their corresponding design vectors containing the values of p^a and p^b exploratory variables. The coefficients e_{ij}^a and e_{ij}^b represent the level-1 residual for the two outcomes. The bivariate model can be presented in the matrix form:

$$\begin{pmatrix} Y_i^a \\ Y_i^b \end{pmatrix} = \begin{pmatrix} X_i^a & 0 \\ 0 & X_i^b \end{pmatrix} \begin{pmatrix} \beta^a \\ \beta^b \end{pmatrix} + \begin{pmatrix} Z_i^a & 0 \\ 0 & Z_i^b \end{pmatrix} \begin{pmatrix} \lambda_i^a \\ \lambda_i^b \end{pmatrix} \qquad (3)$$

Bivariate Model: Random Effects Approach

The outcomes of interest, BMI and NCS are denoted by $Y_{1i}(t)$ and $Y_{2i}(t)$ respectively, for subject i at time t. Assume that the outcomes can be described by two linear mixed models with correlated random effects. A joint model can be constructed from the basic random effect model proposed by Laird and Ware (1982):

$$\begin{aligned} Y_{1i}(t) &= \beta_1(t) + a_{1i} + b_{1i}(t) + \varepsilon_{1i}(t) \\ Y_{2i}(t) &= \beta_2(t) + a_{2i} + b_{2i}(t) + \varepsilon_{2i}(t) \end{aligned} \tag{4}$$

where $\beta_1(t)$ and $\beta_2(t)$ refer to the average changes of the outcomes. The dependent variables (BMI and NCS) are linked together through the joint distribution of their random effects

$$\begin{pmatrix} a_{1i} \\ a_{2i} \\ b_{1i} \\ b_{2i} \end{pmatrix} \sim N(0, D)$$

where D is a matrix of the random effects with the following variance-covariance structure:

$$\begin{pmatrix} \sigma_{a1}^2 & \cdots & \sigma_{a1b2} \\ \vdots & \ddots & \vdots \\ \sigma_{b2a1} & \cdots & \sigma_{b2}^2 \end{pmatrix}$$

The variance-covariance matrix of the model parameter is often derived by maximum likelihood or restricted maximum likelihood method.

The residual components are uncorrelated and independent of the random effects

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right).$$

The variance-covariance structure implies that conditioning on the random effects, both responses are independent. Other cases can be obtained by making additional assumptions about the variance-covariance matrix D. Thiebaut, et al. (2002) outlined further procedures on a number of special cases of the D matrix. For these analyses, models were fitted using the SAS GLIMMIX procedure, while the random effects were introduced through the shared random effects procedures as described by De Gruttola, et al. (1994).

A common objective for joint modeling is to investigate how outcomes are correlated. For example, a researcher might be interested in how the correlation of BMI and NCS are associated with age, while controlling for race and gender. The correlation between outcome variables BMI and NCS is derived from the variance- covariance matrix of the random effect and is given by:

$$r_e = \frac{\sigma_{b_1b_2}}{\sqrt{\sigma_{b_1}^2}\sqrt{\sigma_{b_2}^2}}$$

The marginal correlation between BMI and NCS as a function of time t is given by:

$$r_m(t) = \frac{\sigma_{a_1a_2} + t(\sigma_{a_1b_2} + \sigma_{a_2b_1}) + t^2\sigma_{b_1b_2}}{\sqrt{\sigma_{a_1}^2 + 2t\sigma_{a_1b_1} + t^2\sigma_{b_1}^2 + \sigma_1^2}\sqrt{\sigma_2^2 + 2t\sigma_{a_2b_2} + t^2\sigma_{b_2}^2 + \sigma_{a_2}^2}}.$$

When $t = 0$ the equation reduces to

$$r_m(t) = \frac{\sigma_{a_1a_2}}{\sqrt{\sigma_{a_1}^2 + \sigma_1^2}\sqrt{\sigma_2^2 + \sigma_{a_2}^2}}$$

This formula implies that the marginal correlation cannot be higher than the correlation between random intercepts. If the measurement errors of BMI and NCS are smaller, the marginal correlation at $t = 0$ better approximates the correlation between the random intercepts. The covariance parameters of the random effects and the error components determine the shape of the marginal correlation function. Further, as t increases, the marginal correlation converges to the correlation between the random slopes.

Methodology

Simulation Study

Yang, et al. (2007, 2009, 2010) discussed methods to simulate and test for an over-dispersed Poisson distribution. A simple and efficient way to simulate a mixture of bivariate continuous and over-dispersed Poisson distribution using the specified parameters is:

$$\beta^a = (1, 0.2, -0.2, 1)' a = 1, 5$$
$$\beta^b = (1, 0.2, 0.1, 0.1)', \rho = 0.5, \sigma^2 = 1$$
$$\lambda_i^a = \lambda_i^b = 0.12, n = 100, T = 7$$

Data were simulated using a negative binomial distribution for the case where α = 5. A negative binomial may result from a mixture of Poisson distributions with a Gamma distribution of the mean and a specified shape and scale parameters (R – Documentation). The mean to variance ratio was about 1:5. See Table 1, Figure 1 for a table and plot of the mean and variance for a non-dispersed and see Table 2, Figure 2 for a dispersed Poisson distribution. Using equation 4, 5,000 samples were generated; each sample of size 100 represents a mixed Poisson and continuous longitudinal data with 7 time points. Both outcomes were generated as a linear function of 4 predictors (x_i); the predictors were specified as both binary and continuous. The bias of the estimates using different transformations and correlations were approximately equal.

Results

Tables 3 - 5 summarize the results of the simulated dataset; Log likelihood (-2 Log L), AIC, AICC and BIC together with the means estimates, standard errors and p-values of our estimates. The AIC, AICC and BIC as defined in Akaike (1973, 1985), Sakamota, et al. (1986) and Bozgogan (1987) are specified as: $2(-\log g(x, \hat{\theta}) + d)$, $AIC + \dfrac{2(d+1)(d+2)}{n-d-2}$ and $-2\log g(x, \hat{\theta}) + d(\log(n^*))$ respectively.

These definitions of AIC, AICC and BIC imply that smaller AIC, AICC and BIC estimates provide a better fit for the model. Gurka, et al. (2011) showed that the model with independent correlation structure offers a parsimonious model, but may not always be the best model. However, this study shows that – in the case of mixed Poisson and continuous longitudinal data – the independent and the unstructured correlation tend to perform best and have the lowest AIC, AICC and BIC (see Table 3). In the case of a mixed over-dispersed Poisson and continuous clustered data a number of transformations for the outcome variables were examined. First, the case with no transformation on the outcome was considered; this model

Table 1: Mean and Variance for a Non-Dispersed Poisson Distribution

Simulated Data Results (Non-Dispersed Poisson)		
Replicate	Mean of Poisson	Variance of Poisson
1	4.46	4.95
2	4.04	3.75
3	3.98	3.08
4	3.30	2.79
5	4.54	3.56
6	4.18	3.82
7	4.32	3.81

Figure 1: Mean and Variance Plots for a Non-Dispersed Poisson Distribution
Results Obtained from one Replicate (n = 100, q = 7)

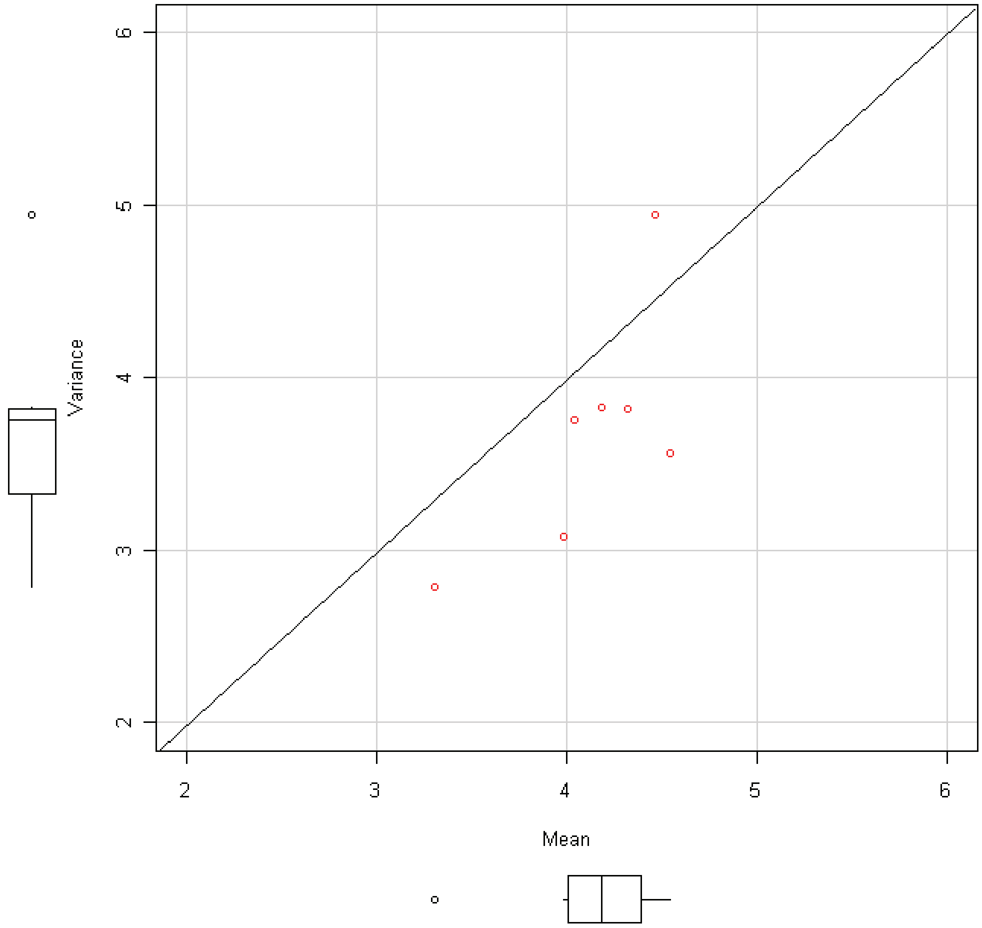
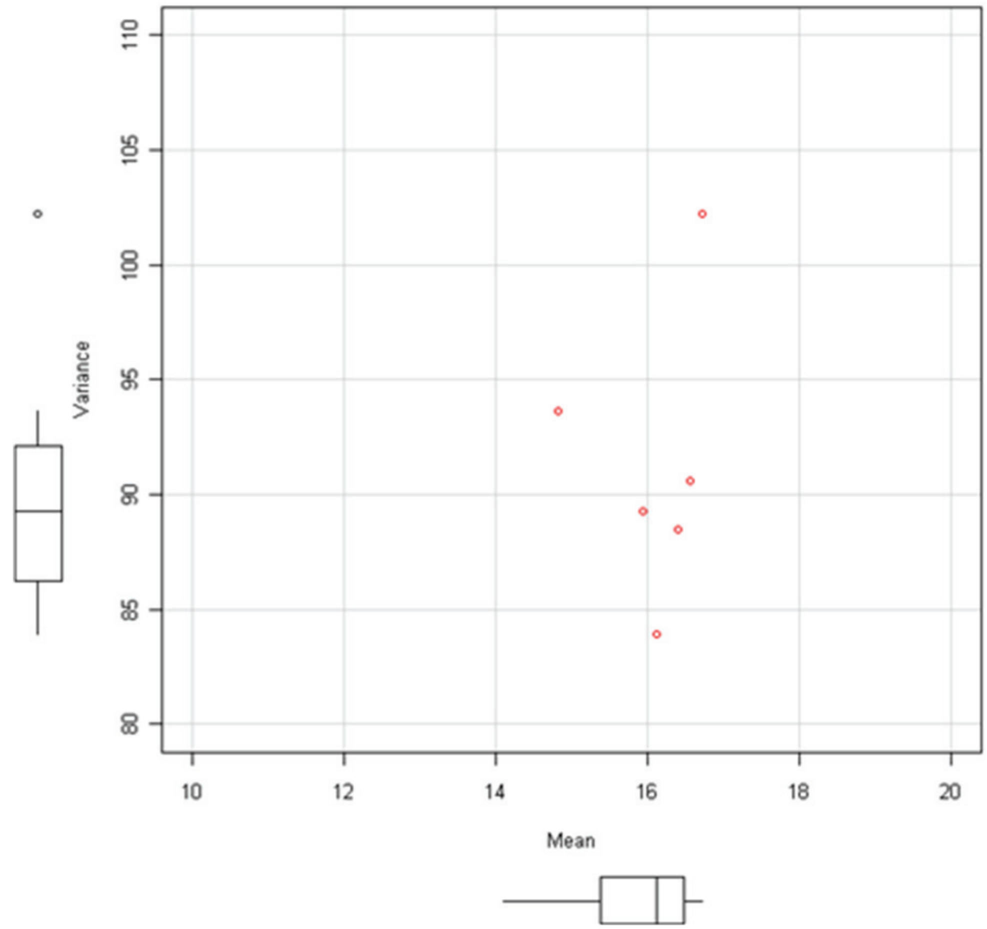


Table 2: Mean and Variance for a Dispersed Poisson Distribution

Simulated Data Results (Dispersed Poisson)		
Replicate	Mean of Poisson	Variance of Poisson
1	15.94	89.24
2	16.56	90.62
3	16.40	88.49
4	16.12	83.94
5	14.10	64.01
6	16.72	102.21
7	14.82	93.62

Figure 2: Mean and Variance plots for a Dispersed Poisson Distribution Results Obtained from one Replicate (n = 100, q = 7)



performed poorly with extremely high values for AIC, AICC and AICC (see Table 4). Next, a number of data transformations were performed and the best model was selected based on the AIC, AICC, BIC and the standard errors of the estimates. The transformation $\log(Y+1)$ performed best for both responses using the independent and the unstructured correlation structure for the data (see Table 5).

Using NHIS data, approximately 264,727 people were interviewed over the 10-year period 1997-2006. The sample size for this analysis was reduced to 42,138 after cleaning. First, BMI values were analyzed with a general linear mixed model. The fixed effects in the model were gender, race and age. The random effects in this model were year (time) and residual (see Table 7). Next, NCS was analyzed with similar fixed effect and random effect variables. The univariate analysis assumed a Poisson error because NCS is count data (see Table 8); however, the resulting fitting algorithm did not converge. The Poisson error might not be a proper residual term for this analysis even though NCS is count data because the means are not equal to the variances (see Table 6). Overdispersion was verified using techniques proposed by Lindsey (1999). Assuming a normal error for NCS, the deviance is 323842.2 which is more than twice the degrees of freedom of 42139, hence the Poisson data is overdispersed.

Table 3: Means, Standard Errors and p-values of Fixed Effects Parameter Estimates

Simulated Data Results (No Overdispersion)					
Variable	Independent	Compound Symmetry	First Order Autoregressive	Toeplitz	Unstructured
-2 Log L	2500.50	2572.83	2572.83	2572.83	2485.63
AIC	2507.60	2580.83	2580.83	2580.83	2507.63
AICC	2507.65	2580.89	2580.89	2580.89	2508.02
BIC	2514.39	2588.48	2588.48	2588.48	2528.66
α	-4.8887	-4.7744	-4.7744	-4.7744	-4.8593
$\beta 1$	27.1193	27.1193	27.1193	27.1193	27.1193
$\beta 2$	0.0034	0.0034	0.0034	0.0034	0.0034
$\beta 3$	0.6846	0.6846	0.6846	0.6846	0.6846
$\beta 4$	-0.0063	-0.0092	-0.0092	-0.0092	-0.0070
Standard Error					
α	1.5987	1.0841	1.0841	1.0841	1.3286
$\beta 1$	0.3131	0.2424	0.2424	0.2424	0.3425
$\beta 2$	0.0274	0.0426	0.0426	0.0426	0.0283
$\beta 3$	0.0423	0.0547	0.0547	0.0547	0.0429
$\beta 4$	0.0394	0.0265	0.0265	0.0265	0.0324
p-value					
α	0.0150	0.0093	0.0093	0.0093	0.0053
$\beta 1$	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
$\beta 2$	0.5084	0.6408	0.6408	0.6408	0.5179
$\beta 3$	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
$\beta 4$	0.5675	0.4547	0.4547	0.4547	0.5820

Table 4: Means, Standard Errors and p-values of Fixed Effects Parameter Estimates

Simulated Data Results (Overdispersed Poisson Without Transformation)					
Variable	Independent	Compound Symmetry	First Order Autoregressive	Toeplitz	Unstructured
-2 Log L	3626.17	3668.83	3668.83	3668.83	3540.68
AIC	3633.57	3676.83	3676.83	3676.83	3562.98
AICC	3633.62	3676.89	3676.89	3676.89	3563.38
BIC	3640.65	3684.48	3684.48	3684.48	3584.30
α	4.8104	5.9763	5.9763	5.9763	6.0410
$\beta 1$	15.7333	15.7333	15.7333	15.7333	15.7333
$\beta 2$	0.0114	0.0114	0.0114	0.0114	0.0114
$\beta 3$	2.8471	2.8471	2.8471	2.8471	2.8471
$\beta 4$	0.0354	0.0061	0.0061	0.0061	0.0045
Standard Error					
α	3.3361	1.7362	1.7362	1.7362	2.1000
$\beta 1$	1.0875	1.0691	1.0691	1.0691	1.1838
$\beta 2$	0.0278	0.0436	0.0436	0.0436	0.0284
$\beta 3$	0.2171	0.2403	0.2403	0.2403	0.1950
$\beta 4$	0.0786	0.0342	0.0342	0.0342	0.0434
p-value					
α	0.2181	0.0308	0.0308	0.0308	0.0357
$\beta 1$	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
$\beta 2$	0.4715	0.6047	0.6047	0.6047	0.4783
$\beta 3$	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
$\beta 4$	0.4694	0.3689	0.3689	0.3689	0.4897

Table 5: Means, Standard Errors and p-values of Fixed Effects Parameter Estimates

Simulated Data Results (Overdispersed Log Transformed Poisson Outcome)					
Variable	Independent	Compound Symmetry	First Order Autoregressive	Toeplitz	Unstructured
-2 Log L	-373.94	-336.83	-336.83	-336.83	-466.84
AIC	-366.74	-330.53	-330.53	-330.53	-443.44
AICC	-366.69	-330.50	-330.50	-330.50	-443.00
BIC	-359.86	-324.51	-324.51	-324.51	-421.07
α	1.9758	1.9803	1.9803	1.9803	1.9286
$\beta 1$	1.1255	1.1242	1.1242	1.1242	1.1809
$\beta 2$	0.0005	0.0005	0.0005	0.0005	0.0005
$\beta 3$	0.1852	0.1850	0.1850	0.1850	0.1953
$\beta 4$	0.0008	0.0007	0.0007	0.0007	0.0005
Standard Error					
α	0.1177	0.1027	0.1027	0.1027	0.1513
$\beta 1$	0.0827	0.0831	0.0831	0.0831	0.1288
$\beta 2$	0.0012	0.0016	0.0016	0.0016	0.0013
$\beta 3$	0.0181	0.0183	0.0183	0.0183	0.0193
$\beta 4$	0.0021	0.0015	0.0015	0.0015	0.0020
p-value					
α	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
$\beta 1$	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
$\beta 2$	0.6150	0.6922	0.6922	0.6922	0.6211
$\beta 3$	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
$\beta 4$	0.6090	0.4955	0.4955	0.4955	0.6082

BMI and NCS were analyzed using a joint bivariate model (see Table 8) with normal and Poisson residuals for the two outcome variables, respectively. The joint bivariate model appears to converge; however, the corresponding univariate analyses for the over-dispersed Poisson error did not converge. This bivariate model is inappropriate because this analysis requires proper consideration of the correlation of the outcomes, which might not be true. In this case, the result of the joint bivariate analysis is similar to the univariate analysis of BMI as the outcome. This may be attributed to the correlation between the outcome variables BMI and cigarettes smoked not being accurately taken into consideration. In order to be sure that the model properly accounts for the association between both outcomes, it is important to ensure that the model assumptions of the independent univariate models are satisfied. For example, if a Poisson error for the univariate analysis of NCS as a dependent variable is assumed, the resulting model might be subject to over-dispersion (see Tables 2 & 6). Yang, et al. (2009) discussed the method of testing and how to manage these types of situations. In order to deal with over-dispersion, NCS is transformed using log(Y+1), where Y is NCS. This model is very similar to the model in equation 1 with $y_{ij}^{a} = \log(y_{ij}^{a} + 1)$.

After transforming to the count variable, NCS, the joint bivariate random effects model is fit. Fixed effects were gender, race, age, age interaction with age and the distribution (normal). The random effects were year and residual (see Table 9); this result is different from the previous joint bivariate model. Furthermore, the analysis of residual from the joint model with normal error for BMI and Poisson error for NCS shows that the model is not performing well. The Poisson model plot of residual is expected to have a funnel shape, but as Figure 3 shows, this is not the case. The model with normal error for BMI and normal error for log-transformed NCS tend to have improved residual plots, based on the distribution of the residuals. This model is also more efficient (see Figure 4) with −2logRe = 579056.2, compared to the model with untransformed NCS (see Figure 3) with −2logRe = 663787.1.

Conclusion

A number of methods have been discussed to handle correlated data with bivariate outcomes (Song, et al., 2008), Yang, et al., (2006, 2007, 2009), but little work has been done in cases with mixed over-dispersed Poisson and clustered continuous outcomes. Yang, et al., (2006, 2009) discussed estimation procedures for bivariate models with both complete and incomplete cases. Yang, et al., (2007, 2009) introduced various methods to test for over-dispersion using a univariate repeated measure data. Fieuws and Verbeke (2004) compared the univariate to the bivariate model with and without correlated random effects using the Hearing Data collected in the Baltimore Longitudinal Study of Aging (BLSA); they concluded that the bivariate correlated model performed best.

The purpose of this study was not to develop a new estimation technique, but rather to show that a simple transformation together with the correct correlation will stabilize the model. Modeling bivariate outcomes is essential when there is an association between primary outcomes or when the question of interest focuses on the joint behavior of multiple outcomes. The extent to which the dependent structure would be considered, however, depends on the question at hand. If interest is primarily on the population response means and the impact of covariates on these means, then a detailed consideration of the transformation and correlation mechanism may not be of significant importance. However, loss of efficiency could result if the assumed working correlation is far from the true correlation (Gardiner, 2009). On the other hand, if there is interest in both marginal and subject-specific inferences (for example, in estimating the growth trajectories of individuals (Potthoff & Roy, 1964), a careful evaluation of the transformation mechanism and correlation structure is of tremendous importance.

The linear mixed model can be used for both marginal and subject specific inference. For example, on the subject-specific inference mean and the population mean, the significance test for this approach depends highly on the chosen covariance structure; therefore, careful consideration in choosing the right correlation structure is required.

Table 6: Number Cigarettes per Day

NHIS Data Results			
Survey Year	N	Mean	Variance
1997	5578	16.27	146.87
1998	4619	15.98	141.91
1999	4316	15.39	144.02
2000	4395	15.32	149.78
2001	4616	14.96	134.03
2002	4158	14.73	128.41
2003	3900	14.33	131.44
2004	3818	14.36	129.28
2005	3840	13.85	120.16
2006	2899	14.00	144.20

Table 7: Modeling BMI with Normal Error

NHIS Data Results					
Effect	Estimate	Standard Error	DF	T-value	P-Value
Intercept	20.75	0.2043	9	101.56	<.0001
Male vs. Female	0.63	0.0366	84263	17.30	<.0001
White vs. Others	–0.23	0.0825	84263	–2.73	0.0064
Black vs. Others	1.20	0.0944	84263	12.67	0.0013
Age	0.25	0.0068	84263	37.13	<.0001
Age*Age	>-0.01	0.0001	84263	–37.22	<.0001

*The estimate of the Age interaction is –0.003

Table 8: Bivariate Model with BMI and Number of Cigarette Smoked per Day

NHIS Data Results						
Effect	Distribution	Estimate	Standard Error	DF	T-value	p-Value
Distribution	Normal	24.62	0.0347	84262	709.21	<.0001
Distribution	Poisson	1.18	0.0229	84262	51.30	<.0001
Male vs. Female		0.17	0.0025	84262	68.05	<.0001
White vs. Others		0.39	0.0067	84262	58.72	<.0001
Black vs. Others		0.08	0.0076	84262	10.56	<.0001
Age		0.04	0.0005	84262	89.48	<.0001
Age*Age		> –0.01	0.0001	84262	–71.98	<.0001

*The estimate of the Age interaction is –0.004

Table 9: Transformed Bivariate Model BMI and Number of Cigarette Smoked Per Day

NHIS Data Results						
Effect	Distribution	Estimate	Standard Error	DF	T-value	p-Value
Distribution	Normal	10.86	0.3838	84263	28.28	<.0001
Male vs. Female		0.39	0.0857	84263	4.510	<.0001
White vs. Others		0.10	0.1933	84263	0.540	0.5903
Black vs. Others		0.68	0.2213	84263	3.050	0.0023
Age		0.15	0.0159	84263	9.25	<.0001
Age*Age*		> –0.01	0.0002	84263	–8.98	<.0001

*The estimate of the Age interaction is –0.0016

Figure 3: Analysis of Residuals: Number of Cigarettes Smoked and BMI

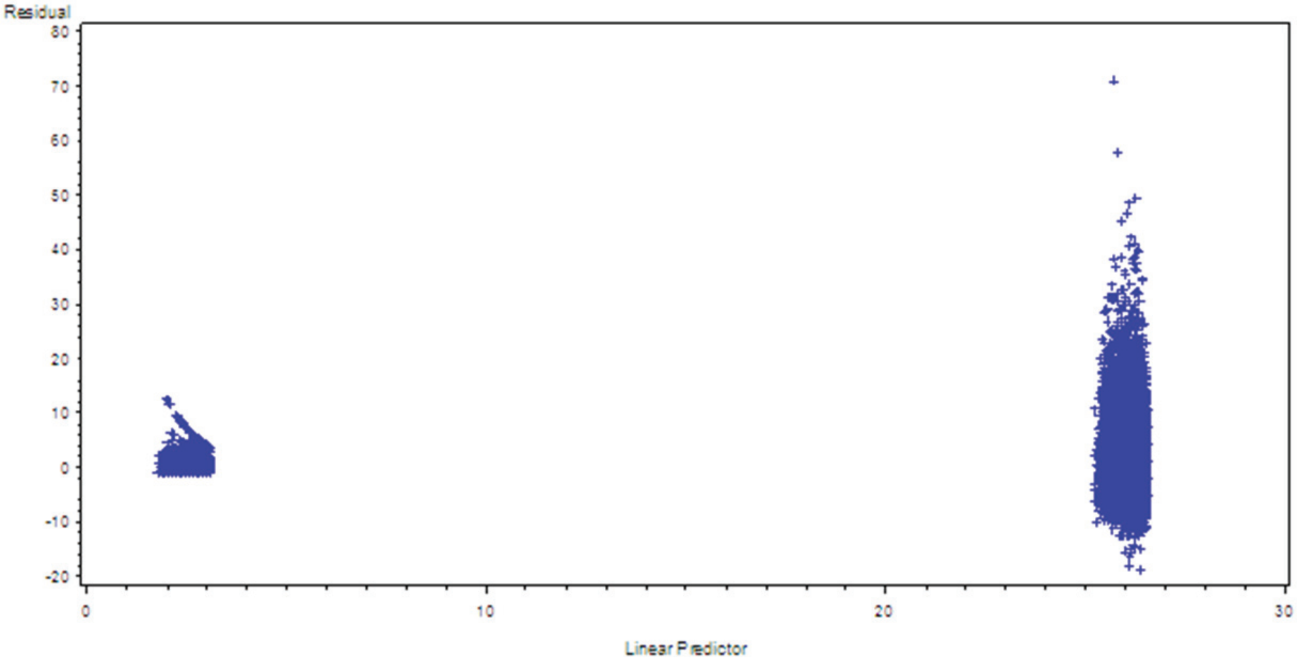
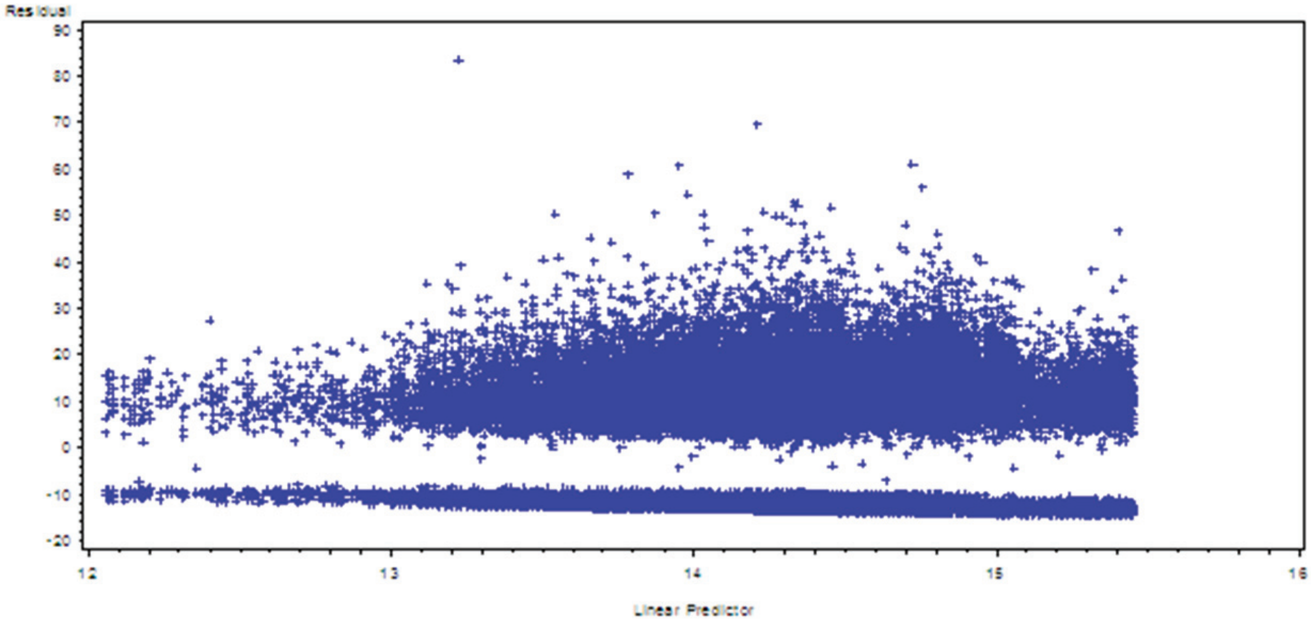


Figure 4: Analysis of Residuals; Transformed Number of Cigarettes Smoked and BMI



The fixed effects estimates with different covariance structures may yield the same values, even though the standard errors of these estimates can vary widely. One objective of data analysis using a linear mixed model is to define an adequate error covariance structure in order to obtain efficient estimates of regression parameters; however, to properly estimate the covariance structure, the normality assumption of the random effect must be met. After both conditions are met, multilevel models are most suitable for analysis of longitudinal data and data with hierarchical structure. (Bock, 1989; Bryk & Raudenbush, 1996; Goldstein, 2003: Hoeksma & Knol, 2001; Raudenbush, 1989; Snijders, 1996).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, B. N. Petrov and F. Csaki (Eds.), 267-281. Budapest: Akademia Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(*6*), 716-723.

Akaike, H. (1985). .Prediction and entropy. In *A Celebration of Statistics*, A. C. Atkinson and S. E. Fienberg (Eds.), 1-24. New York: Springer.

Atem, F. D., Sharma, R. K., & Anderson, S. J. (2010). Fitting Bivariate Multilevel Models To Assess Long Term Changes in Body Mass Index (BMI) and Cigarette Smoking. *Journal of Applied Statistics*, *38*(*9*), 1819-1831.

Bagiella, E., Sloan, R. P., & Heitjan, D. F. (2000). Mixed-effects models in psychophysiology. *Psychophysiology*, *37*, 13-20.

Bellamy, N. (1995). WOMAC osteoarthritis user's guide. *Victoria Hospital Journal of Rheumatology Supplement*, *43*, 49-51.

Bock, R. D. (1989). *Multilevel analysis of educational data*. San Diego, CA: Academic Press.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345-370.

Bryk, A. S., & Raudenbush, S. W. (1987). *Application of hierarchical linear models to assessing change*. *Psychological Bulletin*, *101*(*1*), 147-158.

Bryk, A., & Raudenbush, S. W. (1992). *Hierarchical Linear Models for Social and Behavioral Research: Applications and Data Analysis Methods*. Newbury Park, CA: Sage.

Dempster, A. P., Selwyn, M. S., Patel, C. P., & Roth, A. J. (1984). *Statistical and computational aspects of mixed model analysis*. *Applied Statistics*, *33*, 203-214.

Dmitrienko, A., Tamhane, A. C., & Bretz, F. (2009). Multiple problems in Pharmaceutical Statistics. New York, NY: Chapman and Hall/CRC Press.

Fieuws, S., & Verbeke, G. (2004). Joint modeling of multivariate longitudinal profiles: pitfalls of the random-effect approach. *Statistics in Medicine*, *23*, 3093-3104.

Gardiner, J. C., Lu, Z., & Roman, L. A. (2009). Fixed effects, random effects and GEE: What are the differences? *Statistics in Medicine*, *28*, 221-239.

Goldstein, H. (2003). *Multilevel statistical models*, *3rd Ed*. London: Oxford University Press.

Gomez, B. E., et al. (2005). Covariance Structure is Selected Using AIC and BIC. *Communication in Statistics Simulation and Computation*, 34, 377-392.

Gurka, M. J, Edwards, L. J., & Muller, K. E. (2011). Avoiding bias in mixed model inference for fixed effects. *Statistics in Medicine*, *30*(*22*), 2696-2707.

Hoeksma, J. B., & Knol, D. L. (2001). Testing predictive developmental hypothesis. *Multivariate Behavioral Research*, *36*, 227-248

Karlis, D., & Meligkotsidou, L. (2005). Multivariate Poisson regression with covariance structure. *Statistical Computations*, *15*, 255-265.

Laird, N. M., &Ware, J. H. (1982). *Random-effects models for longitudinal data*. *Biometric*, *38*, 963-974.

Lindsey, J. K. (1999). *Models for Repeated Measurements*, *2nd Ed.* Oxford: Oxford University Press.

McCulloch, C. (2008). Joint modeling of mixed outcome types using latent variables. *Statistical methods in Medical Research*, *17*, 53-73.

Pantazis, N., & Touloumi, G. (2007). Fitting bivariate models for longitudinal data with informative drop-outs using MLwiN. *Multilevel Modeling Newsletter*, *18*, 10-18.

Pawitan, Y., & Self, S. (1993). Modeling disease marker processes in AIDS. *Journal of the American Statistical Association*, *88*, 719-726.

Rao, C. R. (1965). *The theory of least squares when parameters are stochastic and its application to the analysis of growth curves*. *Biometrika*, *52*, 447-458.

Rao, C. R. (1973). *Linear Statistical Inference and it Application*, *2nd Ed.* New York, NY: Wiley.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1886). *Akaike information criterion statistics*. Toyko, Japan: KTK Scientific Publisher.

Snijders, T. (1996). Analysis of longitudinal data using the hierarchical linear model. *Quality and Quantity*, *30*, 405-426.

Song, X., Lee, S., & Hser, Y. (2008). A two-level structural equation model approach for analyzing multivariate longitudinal responses. *Statistics in Medicine*, *27*, 3017-3041.

Thiebaut, R., et al. (2002). Bivariate linear mixed models using SAS proc MIXED. *Computer Methods Programs Biomed*, *69*, 249-256.

Tsiatis, A. A, Degruttola, V., & Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error: Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, *90*, 27-37.

Yang, Y., &Kang, J. (2010). Joint analysis of mixed Poisson and continuous longitudinal data with nonignorable missing values. *Computational Statistics and Data Analysis*, *54*, 193-207.

Yang, Y., Kang, J., Mao, K., & Zhang, J. (2007). Regression models for mixed Poisson and continuous longitudinal data. *Statistics in Medicine*, *26*, 3782-3800.

Yang, Z., Hardin, J. W., & Addy, C. L. (2009). A score test for overdispersion in Poisson regression based on the generalized Poisson-2 model. *Journal of Statistical Planning and Inference*, *139*, 1514-1521.

Yang, Z., Hardin, J. W., Addy, C. L., & Vuong, Q. H. (2007). Approaches for overdispersion in Poisson regression versus the generalized Poisson model. *Biometrical Journal*, *49*(*4*), 565-584.