Journal of Modern Applied Statistical Methods

Volume 16 | Issue 1 Article 28

5-2017

Multivariate Multilevel Modeling of Age Related Diseases

Kapuruge N. O. Ranathunga
University of Colombo, Sri Lanka, nishikaoshadini@gmail.com

Roshini Sooriyarachchi
University of Colombo, Sri Lanka, roshinis@hotmail.com

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 498-517.
doi: 10.22237/jmasm/1493598540

ISSN 1538 − 9472

Multivariate Multilevel Modeling of Age Related Diseases

Kapuruge N. O. Ranathunga
University of Colombo
Colombo, Sri Lanka

Roshini Sooriyarachchi
University of Colombo
Colombo, Sri Lanka

The emerging role of modeling multivariate multilevel data in the context of analyzing the risk factors are examined for the severity of cardiovascular disease diabetes, and chronic respiratory conditions. The modeling phase results leads to some important interaction terms between blood glucose, blood pressure, obesity, smoking and alcohol to the mortality rates.

Keywords: multivariate multilevel model, probit regression, cardiovascular disease and diabetes, chronic respiratory conditions, markov chain Monte Carlo

Introduction

Aging increases susceptibility to age-associated diseases and some of these diseases may increase mortality among adults worldwide. The focus of this study is on cardiovascular disease and diabetes (CDD) and chronic respiratory conditions (CRC). These are life-threatening diseases with increasing incidence. Also, there is a geographical effect of the mortality rates of these diseases (World Health Organization, 2005). Therefore, countries are grouped into continents geographically, but vary across continents. This establishes the need of multilevel hierarchical analysis

Because the existence of a high correlation between these variables, and the presence of some common risk factors to these related diseases, a multivariate multilevel concept was used to identify the joint effects of some risk factors on these two diseases to analyze data more appropriately. A multivariate multilevel model can be considered as a collection of multiple dependent variables in a hierarchical nature. When the effect of a set of explanatory variables on a set of dependent variables shows a considerable difference then it can be handled only by means of a multivariate analysis (Snijders & Bosker, 2012).

Kapuruge N. O. Ranathunga is a former Assistant Lecturer in the Department of Statistics. Email her at nishikaoshadini@gmail.com.

Table 1. Description of the data and its abbreviations

Variable Name	Identifier	Category	Code
Cardiovascular Diseases and Diabetes (per 100,000 population)	CDD	<220	1
		220-370	2
		>370	3
Chronic Respiratory Conditions (per 100,000 population)	CRC	<20	1
		20-50	2
		>50	3
Population using improved drinking water sources (%)-2011[a]	Water	<88	1
		88-98	2
		>98	3
Population using improved sanitation (%)-2011[a]	Sanitation	<40	1
		40-80	2
		>80	3
Population using solid fuels (%)-2011[a]	Solid_Fuel	<20	1
		20-70	2
		>70	3
Prevalence of raised fasting blood Glucose among adults aged ≥ 25 years (%)-2008[a]	B_Glucose	<7.5	1
		7.5-11.5	2
		>11.5	3
Prevalence of raised blood pressure among adults aged ≥ 25 years (%)-2008[a]	B_Pressure	<25	1
		25-35	2
		>35	3
Adults aged ≥ 20 years who are obese (%)-2008[a]	Obese	<13	1
		13-24	2
		>24	3
Alcohol consumption among adults aged ≥ 15 years (litres of pure alcohol per person per year)-2008[a]	Alcohol	<4	1
		4-10	2
		10-16	3
		>16	4
Prevalence of smoking any tobacco product among adults aged ≥ 15 years (%)-2009[a]	Smoking	<12	1
		12-24	2
		24-36	3
		>36	4

Note: *a*) country level (1st level) variables

Considered here is a multivariate multilevel analysis approach by using Bayesian methods. Data for this study were obtained from the World Health Organization (2013). The dataset consists of worldwide mortality rates among

adults aged 30-70 years. Due to the incompleteness of the records, a multiple imputation (MI) was conducted to variables Smoking, Water and Sanitation prior to fitting the models (Sterne et al., 2009). The MI procedure requires the variables to be imputed to be normally distributed or categorical. Water and Sanitation did not follow a normal distribution, and were categorized to perform the MI (Table 1), and are considered ordinal categorical variables for the modeling.

Given in Table 1 are the variables and their respective categories with abbreviations. The continuous data were discretized in to ⅓ splits based on percentiles to obtain respective categories. Alcohol and smoking were categorized into ¼ splits to obtain more explicate categories due to the expansion of the data.

Methodology

Univariate analysis using Zhang and Boos test

Before carrying out the modeling it is essential to determine the nature of the strength of the relationships between explanatory variables and response variables. However due to the natural hierarchy of the observations, Zhang & Boos (1997) developed the Generalized Cochran Mantel Haenszel (GCMH) test. There are three different types of test statistics proposed by Zhang and Boos. These are T_{EL}, T_P and T_U. From these, T_P is preferable to T_U and T_{EL} (Jayawardana and Sooriyarachchi, 2014). Simulation studies showed it maintains error values even for a small number of strata (Zhang and Boos, 1997).

Structure of the Multivariate Multilevel Probit Regression Model

Although the logit link is the most common, the multivariate model for binary responses was developed for the probit link in MLwiN 2.10 (Rasbash et al., 2009). Due to the unavailability of a proper documentation of the theory regarding multivariate multilevel binary probit models, it was discussed based on the theory regarding multivariate multilevel probit models for the ordered categorical responses, given by Grilli and Rampichini (2003).

Simple Probit Regression Model Suppose the response of interest which is known as Y can take values 1 and 0 where 1 = higher risk, 0 = lower risk and x can be denoted as set of explanatory variables.

$$\Pr(Y=1|x)=F\left(\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_kx_k\right), \quad (1)$$

where $F(.)$ is a function such that $F: x \rightarrow [0, 1]$, for all x that belongs to the real line.

The probit model assumes that the function $F(.)$ follows a Normal (cumulative) distribution,

$$F(x) = \Phi(x) = \int_{-\infty}^{x} \varphi(z)\, dz\,, \tag{2}$$

where, $\varphi(z)$ is the Standard Normal Density Function.

$$\varphi(z) = \frac{\exp\left(-\frac{z^2}{2}\right)}{\sqrt{2\pi}} \tag{3}$$

Multivariate Multilevel Probit Regression Model Let $Y_{ij}^{(h)}$ be the h^{th} ($h = 1, 2, \ldots, H$) observed binary variable for the i^{th} ($i = 1, 2, \ldots, I$) observation of the j^{th} ($j = 1, 2, \ldots, J$) cluster. Assume that each of the observed responses $Y_{ij}^{(h)}$, which takes values in {1, 2} (for the sake of simplicity, assume it as C) is generated by a latent variable $\tilde{Y}_{ij}^{(h)}$ through the following relationship:

$$\left\{Y_{ij}^{(h)} = C^{(h)}\right\} \text{ if and only if } \left\{\gamma_{C^{(h)}-1}^{(h)} < \tilde{Y}_{ij}^{(h)} < \gamma_{C^{(h)}}^{(h)}\right\} \tag{4}$$

where the threshold satisfies $-\infty = \gamma_0^{(h)} \leq \gamma_1^{(h)} = +\infty$ and γ represents the corresponding value of the response variable when h and c takes values as in equation (4).

Now, consider the following multivariate two-level null model for the latent variables:

$$\tilde{Y}_{ij}^{(h)} = \alpha^{(h)} + u_j^{(h)} + \varepsilon_{ij}^{(h)}, h = 1,..., H\,, \tag{5}$$

where for each h, $\alpha^{(h)}$ is the mean, $u_j^{(h)}$ is the cluster's random effect (level two error) and $\varepsilon_{ij}^{(h)}$ is the individual's disturbance (level one error). The errors are assumed to be distributed as

$$\left[\varepsilon_{ij}^{(1)},...,\varepsilon_{ij}^{(H)}\right]' \sim iidN\left(0,\Sigma_{\epsilon}\right) \text{ and } \left[u_j^{(1)},...,u_j^{(H)}\right]' \sim iidN\left(0,\Omega\right) \quad (6)$$

For example, for $H = 2$ the covariance matrices are,

$$\Sigma_{\epsilon} = \begin{pmatrix} \sigma_{\epsilon 1}^2 & \sigma_{\epsilon 1\epsilon 2} \\ \sigma_{\epsilon 1\epsilon 2} & \sigma_{\epsilon 2}^2 \end{pmatrix}, \Omega = \begin{pmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{pmatrix} \quad (7)$$

Moreover, the first and second level errors are assumed to be independent. The previous model specification implies the following conditional covariance structure for any couple of latent variables $\left[\tilde{Y}_{ij}^{(h)},\tilde{Y}_{ij}^{(k)}\right]$:

$$Cov\left[\tilde{Y}_{ij}^{(h)},\tilde{Y}_{i'j'}^{(k)} \mid u_j^{(h)},u_{j'}^{(k)}\right] = E\left[\varepsilon_{ij}^{(h)}\varepsilon_{i'j'}^{(k)}\right] = \left\{ \begin{array}{ll} \sigma_{\epsilon_h}^2 & \text{if } k = h, j = j', i = i' \\ \sigma_{\varepsilon_h\varepsilon_k} & \text{if } k \neq h, j = j', i = i' \\ 0 & \text{otherwise} \end{array} \right\}. \quad (8)$$

The unconditional covariance structure is

$$Cov\left[\tilde{Y}_{ij}^{(h)},\tilde{Y}_{i'j'}^{(k)}\right] = E\left[\varepsilon_{ij}^{(h)}\varepsilon_{i'j'}^{(k)}\right] + E\left[u_j^{(h)}u_{j'}^{(k)}\right], \quad (9)$$

with $Cov\left[\tilde{Y}_{ij}^{(h)},\tilde{Y}_{i'j'}^{(k)}\right] = 0$ if $j \neq j'$.

The correlation between the same variable for two distinct individuals of the same cluster, called the Intra Cluster Correlation Coefficient (ICC), is stated below.

$$Corr\left[\tilde{Y}_{ij}^{(h)},\tilde{Y}_{i'j}^{(k)}\right] = \tau_h^2\left(\sigma_{\epsilon_h}^2 + \tau_h^2\right), h = 1,...,H. \quad (10)$$

ICC also represents the proportion of variance explained by the clusters.

Variable selection and model comparison

Consider a subset of covariate and cofactors from the pre-specified set of variables, which best describes the dependent variables. A Forward Selection

procedure along with the Wald Statistic is specifically used for this purpose. MLwiN 2.10 does not use maximum likelihood estimation for estimating the parameters because it is computationally difficult. Therefore as a solution to that, the quasi-likelihood methods were implemented. This shows the inability of considering usual likelihood ratio tests for comparing models.

These methods are implemented by transforming discrete response model to a continuous response model based on a Taylor series expansion. Then, the model becomes linear and then estimation is carried out using Iterative Generalized Least Squares (IGLS) or Reweighted IGLS (RIGLS). These transformations require an approximation known as Marginal Quasi-Likelihood (MQL) and Predictive Quasi-Likelihood (PQL) and can be comprised with Taylor series expansions of either first order terms or second order terms. However, when the sample sizes within level 2 units are small, the first order MQL procedures may lead to biased estimates (Rasbash et al., 2009). Therefore, the second order PQL procedure was adopted.

However due, to some convergence and stability problems it was followed by a Markov Chain Monte Carlo (MCMC) method, which is an alternative to likelihood based estimation procedure.

Parameter Interpretation for a multivariate multilevel model with a probit link function

Interpretation of the coefficients in probit regression is not as straightforward as other regressions. The increase in probability for a unit increment in a given predictor depends on both the values of the other predictors and the initial value of the given predictors. Because the final model presents a lot of interactions and deals with multivariate data in a hierarchical nature, this procedure is more complex and time consuming. Therefore, the probability differences for unit increase of covariates when the other continuous covariates are at their average levels and the categorical covariates are at their base level were considered. Due to the inconvenience of calculating the corresponding probability differences in manual form, a SAS program was used for this purpose.

Residual Analysis and Model Adequacy

It is essential to assess the appropriateness of the fitted model by evaluating the adequacy. Because handling data in a multivariate multilevel framework is a novel approach, diagnostic techniques specifically designed for this scenario are less available. Though the specifications of the models are different according to

the types of variables, the methods of residual analysis and model adequacy are common to all models in a hierarchical nature. Rasbash et al. (2009) presented the theory regarding a basis model having a continuous response in a multilevel data.

Multiple Imputation

The original dataset has small number of observations, removing the records with missing data may cause to create a rather small dataset and it leads to exclude approximately 38% of records. This would cause biased results, because there is a high chance of excluding the low- and middle-income countries from the analysis due to the unavailability of proper information systems. Furthermore, the Missing Data Mechanism (MDM) of this dataset takes the form of Missing At Random (MAR) (Rubin, 1976) and it happens when the missingness depends on a specific variable, but not the value of the variable including missing data (Howell, 2012). In the current dataset, low- and middle-income countries might be less inclined to report their health information due to the unavailability of proper information systems. Therefore, the probability of reported health status is unrelated to the level of health within these low- and middle-income countries and hence the data can be considered MAR. Accordingly, MI was carried to this dataset by using REALCOM (Carpenter et al., 2011) software.

Results

There are three hierarchical levels. Level one consists of the multivariate structure. Level two consists of countries and level three consists of continents. There are 10 variables in the dataset; the countries are clustered within continents. The dataset consist of two response variables termed as CDD and CRC and eight continuous explanatory variables at the country level: Water, Sanitation, Solid_Fuel, B_Glucose, B_Pressure, Obese, Alcohol and Smoking. To implement the univariate analysis these variables were discretized. Although originally there were 195 countries in the dataset, after removing some observations with many missing values and then performing imputation techniques to the variables Smoking, Water and Sanitation, that number was reduced to 186.

Compiled in Table 2 are the *p*-values of the univariate analysis for the associations between imputed explanatory variables with the two outcome variables and the composite variable before and after imputation. The test carried out was the GCMH test. Continent to which the countries belong is used as the second level variable to stratify data accordingly.

Table 2. T_P statistic test results for imputed variables with the responses.

Disease	Explanatory variable	Before			After		
		T_P	DF	*p*-value	T_P	DF	*p*-value
CDD	*Water*	21.888	4	0.00021	23.669	4	9.30E-05
	Sanitation	8.250	4	0.08300	9.954	4	0.04120
	Smoking	8.791	6	0.18600	7.244	6	0.29900
CRC	*Water*	21.302	4	0.00028	21.752	4	0.00022
	Sanitation	23.544	4	9.85E-05	25.307	4	4.36E-05
	Smoking	9.805	6	0.13300	11.118	6	0.08500
CDC+CRC	*Water*	27.201	6	0.00013	27.422	6	0.00012
	Sanitation	19.810	6	0.00299	21.633	6	0.00141
	Smoking	16.869	9	0.05080	11.812	9	0.22410

Note: Consider 20% level of significance

As noted in Table 2, the variables that were considered to be insignificant before imputation remained to be so while those that were significant before imputation remained to be significant apart from the variable Smoking coming under CDD and CDD+CRC which was significant before imputation but had become insignificant after imputation.

Univariate analysis for identifying country level factor impact on the response

Because of the stratified nature of the data, GCMH test was used with a liberal significance level of 20% as explained in Collett (1991). This significance level can be increased because more severe significance levels can lead to the exclusion of potentially useful predictor variables. The requisite calculations were performed using the R-macro developed by De Silva and Sooriyarachchi (2012). Prior to implementing GCMH test, the correlation between CDD and CRC was identified using Pearson's correlation test. For that, two diseases were taken as their continuous form. According to the correlation matrix, there is a significant positive correlation (0.680) that exists between CDD and CRC. Therefore, it is appropriate to perform the Multivariate Multilevel analysis on CDD and CRC.

As noted in Table 3, the two diseases were split into binary outcomes in order to maintain the simplicity of the analysis. Otherwise the resulting composite outcome might have large number of categories and it would be more complex to proceed. The categorization was done by considering the cut-points of worldwide

mortality rates for the two diseases together with the aid of specialists in the field of medicine (World Life Expectancy, n.d.).

Table 3. Categorization of the diseases and description of combined levels

Code (Category)		**Coding for Composite outcomes**
CDD	**CRC**	
1 (<300)	1 (<30)	1
1 (<300)	2 (≥30)	2
2 (≥300)	1 (<30)	3
2 (≥300)	2 (≥30)	4

Compiled in Table 4 are the results of the univariate test, which was carried out to check the significance of country level covariates in the presence of continent as the respective stratification factor for the composite outcome of CDD and CRC.

Table 4. Test Results for composite variable of two diseases vs. Risk Factors

Risk Factors	T_P	**DF**	***p*-value**
Water	27.201	6	0.00013
Sanitation	19.810	6	0.00299
Solid_Fuel	31.403	6	2.10E-05
B_Glucose	14.572	6	0.02390
B_Pressure	21.195	6	0.00170
Obese	19.385	6	0.00356
Alcohol	15.797	9	0.07120
Smoking	16.869	9	0.05080

All the risk factors are significant at a liberal 20% level and the variable Solid Fuel shows the most significance. It implies that there is a higher tendency of getting the disease for the people who are using solid fuel for their day-to-day work.

Fitting a multivariate multilevel probit regression model

Before applying the modeling techniques two diseases were categorized into binary splits as in Table 3. Water and Sanitation were taken as ordered categorical variables while others were taken as their original continuous form. For the multivariate multilevel analysis, there are two types of parameter estimates named

as separate coefficients and common coefficients. Due to that, the model building procedure would be more complex and cumbersome. Therefore, several methods were adopted for the simplification and to obtain an adequate model. For the estimation, the 1st order MQL method was followed by the 2nd order PQL method. It was again followed by the MCMC method to obtain Wald statistic values.

Results indicated that improved drinking water sources and improved sanitation may lead to decrease the incidence of both diseases. This means that the incidence of diseases is increasing when the quality level of water and sanitation are decreasing. Therefore, it would be more meaningful and practicable to get the highest level as the reference for both water and sanitation. Presented in Table 5 are the cofactors and their respective base categories used in the modeling phase.

Table 5. Variables and corresponding base categories

Cofactors	Base category
Water	≥98%
Sanitation	≥80%

At the 1st stage, each factor/covariate was fitted separately and the corresponding Wald statistic value was computed. The *p*-value of the statistic was then compared with the 5% significance level to assess the significance of the coefficient. However, because of the Deviance Information Criteria (DIC) is not available in the MLwiN for the multivariate multilevel scenario, the model building procedure was solely based on the Wald statistic. Forward selection procedure was implemented to identify the main effects. If Wald statistic values for separate coefficients are quite close, the common coefficients should be used as parameter estimates. This argument was used for selecting the other terms as common or separate.

At the 2nd stage each interaction term was fitted separately to the final main effects model. Because there are many interactions pertaining to the variables, fitting all would be more cumbersome and MLwiN would not respond to most of them. Therefore, only the interactions which were significant for the two univariate binomial multilevel logistic regressions for CDD and CRC were considered. However, because there were separate and common coefficients, the interactions were added according to the final main effect model. For an example, consider the B_Glucose*Alcohol interaction. In the final model B_Glucose and Alcohol were fitted as a common coefficient. This means to fit the

B_Glucose*Alcohol interaction also as common coefficients. Figure 1 represents the output of the final interaction model.

$$
\begin{aligned}
&\text{resp}_{1jk} \sim \text{Binomial}(\text{n}_{1jk}, \pi_{1jk})\\
&\text{resp}_{2jk} \sim \text{Binomial}(\text{n}_{2jk}, \pi_{2jk})\\
&\text{resp}^*_{1jk} \sim \text{N}(XB, \Omega)\\
&\text{resp}^*_{2jk} \sim \text{N}(XB, \Omega)\\
&\text{probit}(\pi_{1jk}) = 0.204(0.028)\text{B\_Pressure.CDD}_{ijk} + 0.341(0.450)\text{Sanitation\_1.CDD}_{ijk} +\\
&\qquad 0.407(0.390)\text{Sanitation\_2.CDD}_{ijk} + -0.043(0.023)\text{Obese.CDD}_{ijk} +\\
&\qquad 0.002(0.001)\text{B\_Pressure.Obese.CDD}_{ijk} + -0.087(0.043)\text{Smoking.CDD}_{ijk} +\\
&\qquad 0.003(0.001)\text{B\_Pressure.Smoking.CDD}_{ijk} + u_{0jk}\text{bcons.1}_{ijk} + h_{jk}\\
&\text{probit}(\pi_{2jk}) = 0.232(0.024)\text{B\_Pressure.CRC}_{ijk} + 1.766(0.593)\text{Sanitation\_1.CRC}_{ijk} +\\
&\qquad 1.008(0.387)\text{Sanitation\_2.CRC}_{ijk} + -0.252(0.044)\text{Obese.CRC}_{ijk} +\\
&\qquad 0.008(0.001)\text{B\_Pressure.Obese.CRC}_{ijk} + 0.192(0.046)\text{Smoking.CRC}_{ijk} +\\
&\qquad -0.007(0.001)\text{B\_Pressure.Smoking.CRC}_{ijk} + u_{1jk}\text{bcons.2}_{ijk} + h_{jk}\\
&h_{jk} = \beta_{2k}\text{cons.12} + 1.980(0.462)\text{Water\_1.12}_{jk} + 1.092(0.315)\text{Water\_2.12}_{jk} +\\
&\qquad 0.582(0.039)\text{B\_Glucose.12}_{jk} + -0.387(0.055)\text{Alcohol.12}_{jk} +\\
&\qquad 0.058(0.008)\text{B\_Glucose.Alcohol.12}_{jk} + -0.022(0.001)\text{B\_Glucose.B\_Pressure.12}_{jk} +\\
&\qquad -0.009(0.003)\text{Obese.Alcohol.12}_{jk}\\
&\beta_{2k} = -7.419(0.919) + v_{2k}
\end{aligned}
$$

Figure 1. Final interaction model

According to Figure 1, it can be seen that though there are two levels originally present in the data, the MLwiN is recognized it as three levels. This is because the MLwiN treats the outcomes of two diseases responses as the 1st level (*i*). Therefore $resp_{1jk}$ refers to the number of responses for the disease 1 (CDD) made by the j^{th} country those who are clustered within the continent *k*. Similarly, $resp_{2jk}$ refers to the number of responses for the disease 2 (CRC) made by the j^{th} country those who are clustered within the continent *k*. As a result of that, $resp_{ijk}$ can take either zero or one for all countries in the study. Moreover, n_{1jk} and n_{2jk} always take the value 1, because each country always gives a single response.

Continent level variance component analysis

In order to justify the suitability of applying the multilevel concept, it is advisable to first look at the significance of the continent level variance. This can be checked by the following hypotheses

H_0: Continent level residual variance is zero
H_1: Continent level residual variance is not zero

Because zero is not included in the 95% credible interval (0.254, 5.109), H_0 is rejected and concluded that the continent level variance is significant implying that the multilevel approach for the multivariate context is suitable.

Residual analysis of the final model

After fitting the model the model adequacy was checked. For that purpose, Caterpillar plots and Normal probability plots were used. According to the Caterpillar plot in Figure 2, four residuals do not contain zeros in their 95% confidence bands. These imply significant differences from the overall mean predicted by the fixed part from the model. Moreover, it can be seen that two continents show a negative residual deviation while another two show positive deviations. Therefore it is possible to conclude that these four continents contribute to a high continent effect on the mortality rates of CDD and CRC. These four continents are North America, Europe, Asia and Oceania respectively. Figure 4 illustrates these continent variations more clearly. The continents that have a lower risk are symbolized by green, and higher risk are symbolized by red.

It is suggested in Figure 3 the points are approximately through the 45˚axis indicating that the residuals are approximately normally distributed. However, because of the number of residuals is less, it is hard to conclude the assumption of normality by eye inspection and unable to conduct the Anderson-Darling test with the number of residuals less than seven.

Multivariate Multilevel techniques have recently been developed in the field of statistics and its applications and analysis techniques are very rare. Therefore a suitable goodness of fit test has not yet been developed to evaluate the adequacy of the fitted model. Because there are no other techniques available, the model adequacy was solely dependent on the caterpillar plot and the normal plot.

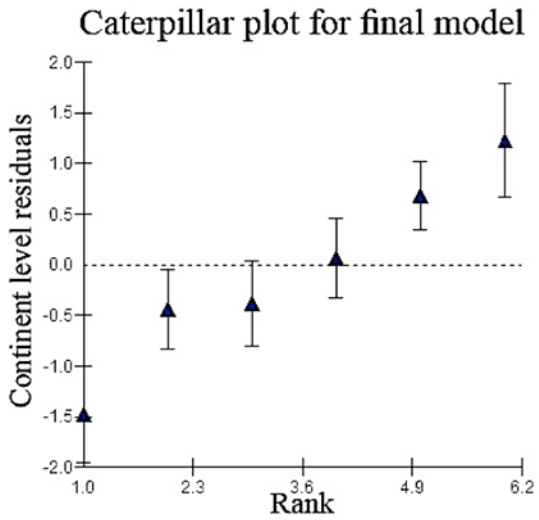

Figure 2. Estimated continent level residuals for the final model

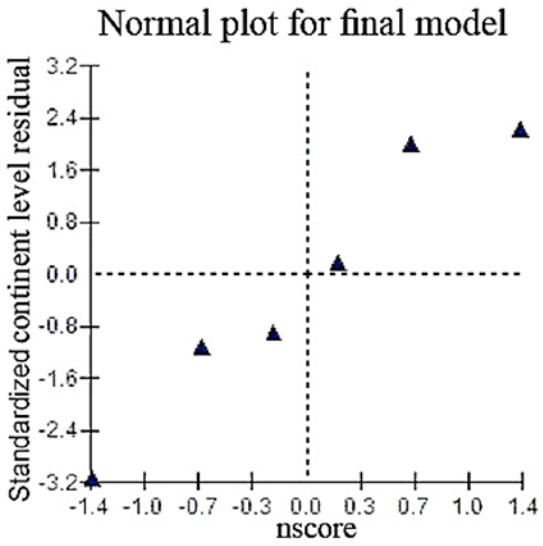

Figure 3. Normal plot for continent level residuals

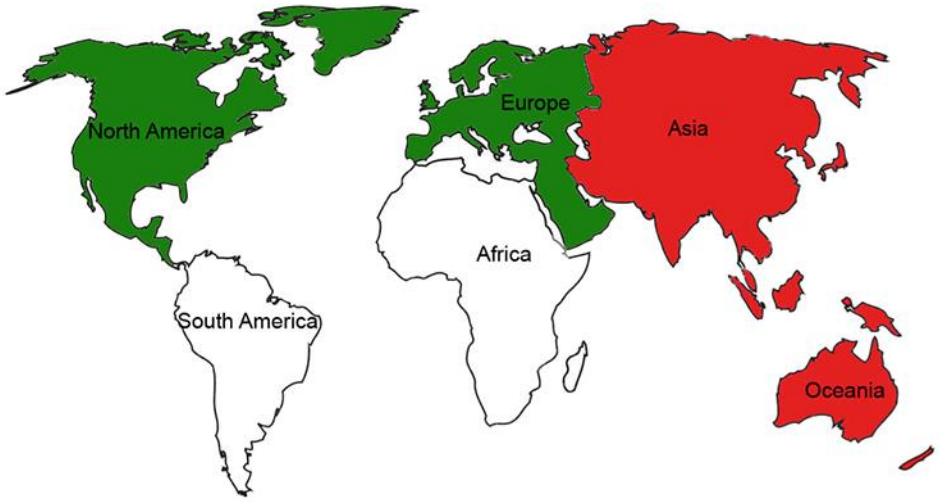

Figure 4. Continent level variations for CDD and CRC

Interpretation and calculation of the parameter estimates

Because the model consists of two equations, due to the multivariate concept this section consists of step-by-step interpretation of each explanatory variable for the two diseases separately. The calculated probability differences are represented in Table 6 and 7.

The results of Table 6 indicate the following important conclusions. The probability of being in the higher group of CDD is 0.6478 higher when Water is at level 1 and 0.3106 higher when Water is at level 2 when compared to level 3 while all the other continuous variables are taken at average and the Sanitation is taken at the base level. However it can be seen that both levels of Sanitation do not have a significant impact for this scenario.

B_Pressure has common interactions with Obese, Smoking and B_Glucose. The probability of CDD being in the higher level compared to the lower level is 0.0149 times more when B_Pressure is increased by one unit and all other variables are at an average and water and sanitation are at base levels.

According to the available medical literature (What Are the Health Risks of Overweight and Obesity?, 2012), it was found that Obesity has shown a higher impact on CDD together with the B_Pressure rather than individually. Therefore, it is more meaningful to identify the combined effect of B_Pressure and Obesity to CDD. The results shows the probability of being in the higher CDD category

compared to the lower CDD category is 0.0081 higher when B_Pressure and Obesity are both increased by one unit when all continuous variables at average and Water and Sanitation are at the base levels.

Table 6. Probability differences for CDD

Term	Probability difference
[a]Water 1	0.6478
[a]Water 2	0.3106
[b]Sanitation 1	Not significant
[b]Sanitation 2	Not significant
[c]B_Pressure = z + 1	0.0149
[c]Obese = y + 1, B_Pressure = z + 1	0.0081
[c]Smoking = s + 1	0.0012
[c]B_Glucose = x + 1	0.0530
[c]Alcohol = w + 1	0.0056

Note: All terms assume continuous variables at average; a) assumes Sanitation = base level; b) assumes Water = base level; c) assumes Sanitation, Water = base level

Table 7. Probability differences for CRC

Term	Probability difference
[a]Water 1	0.5745
[a]Water 2	0.2344
[b]Sanitation 1	0.4911
[b]Sanitation 2	0.2067
[c]B_Pressure = z + 1, Smoking = s + 1	0.0457
[c]Obese = y + 1	-0.0056
[c]B_Glucose = x + 1	0.0341

Note: All terms assume continuous variables at average; a) assumes Sanitation = base level; b) assumes Water = base level; c) assumes Sanitation, Water = base level

Similarly, the probability of CDD being in the higher level compared to the lower level is 0.0012 times higher when Smoking is increased by one unit, 0.053 times more when B_Glucose is increased by one unit and 0.056 times more when Alcohol is increased by one unit while all other variables are at an average and water and sanitation are at base levels

For CRC, the probability of being in the higher group is 0.5745 more when Water is at level 1 and 0.2344 more when it is at level 2 when compared to level 3 while all the other continuous variables are taken at the average and the Sanitation is taken at the base level. Similar to Water, the probability of being in the higher

group of CRC is 0.4911 more when Sanitation is at level 1 and 0.2067 more when Sanitation is at level 2 when compared to level 3. Therefore, it can be seen that when the usage of improved Water and Sanitation sources decreases, the probability of being the higher group of CRC increases.

Smoking has a higher impact to CRC together with B_Pressure rather than individually (Kenny, n.d.). Therefore, when considering the combined effect of those two, the probability of being in the higher CRC category compared to the lower one is 0.0457 times more when both B_Pressure and Smoking are increased by one unit while all continuous variables are at average and Water and Sanitation are at the base levels.

Similarly, the probability of CRC being in the higher level compared to the lower level is 0.0341 times more when B_Glucose is increased by one unit, 0.0693 times more when Alcohol is increased by one unit and 0.0056 times lower when Obesity is increased by one unit while other variables are at an average and water and sanitation are at base levels. Though the latter result seems to be contradictory, it is not so as past medical evidence has suggested that thin people are more prone to get CRC than fat people (Schols et al., 1998).

Discussion

When the usage of unimproved water sources increases, the probability of occurrence of deaths for CDD and CRC also increases. Past evidence also indicated this relationship. Fodor et al. (1973) showed the proportion of mortality rates for CDD was higher in the soft water areas than hard water areas. It was further shown there was a macro geography variation for CDD. Those findings tally with the findings in this study because here also CDD shows a continent level variation. They also showed CRC has an impact from the variable Water. But it is a less known thing. However, officials at the US Environmental Protected Agency suggested heavy rainfall events cause storm water overflow that may contaminate water bodies used for drinking with other bacteria. It may cause to get illnesses, including ear, nose, and throat infections (Climate impacts on Human Health, n.d.).

Although CDD has no impact from Sanitation, the probability of being in the higher group of CRC increases due to the usage of unimproved sanitation sources. When analyzing risk factors for diseases, the focus is less given for the environmental factors such as water, sanitation etc. However, it was shown the usage of unimproved Water and Sanitation sources have more impact to the diseases CDD and CRC. According to Briggs's (2003), unsafe water, poor

sanitation and poor hygiene seem to be one of the major sources of exposure for these types of diseases.

National Heart, Lung and Blood Institute (What Are the Health Risks of Overweight and Obesity?, 2012) claimed most people who have type 2 diabetes are overweight and also it leads to heart failures. Furthermore, they have shown that the chances of having high blood pressure are greater if people are overweight. This joint impact of B_Pressure and Obesity on CDD by showing the probability of occurrence of death in CDD increases when both B_Pressure and Obesity are increased by one unit.

When considering CRC, medical evidence (Kenny, n.d.) suggests that chronic obstructive pulmonary disease (COPD) usually cause by smoking and continuous smoking for a long time causes to increase breathing difficulties and also causes to increase blood pressure. As a result of that it can put a heavy strain on the heart muscle and creates heart failures. After that Respiratory failures occur as the final stage of COPD (Kenny, n.d.). This factor shows that there is an interesting flow by beginning from smoking through the increment of blood pressure to the respiratory failures. This further demonstrates an interesting relationship by showing increase in the probability of being in the higher level of CRC compares to the lower level when B_Pressure and Smoking are both increased by one unit.

Some medical evidence (Schols et al., 1998) suggested it is difficult to identify a suitable relationship between Obesity and CRC. A decrement of the probability being in the higher level of CRC for a unit increment of Obesity was shown. However, a large epidemiologic study showed overweight and obesity in patients with COPD was associated with a decreased risk of death compared with normal weight (Schols et al., 1998). Therefore, it might be concluded that thin people are more prone to get CRC than obese people. Furthermore, North America and Europe show a less risk of having CDD and CRC while Asia and Oceania show a higher risk with CDD varies less with continent while CRC varies more.

Limitations of the study

In the advanced analysis phase, logistic and probit regression models were fitted with the continuous explanatory variables. The models contained common interactions as well as cross interactions. Therefore, it was more complex to obtain corresponding confidence intervals for the odds ratios and for the predicted

probabilities and hence the significance/non-significance of the estimates could not be evaluated.

The interpretations of coefficients in multivariate multilevel binary probit regression models are not as simple as in other models (i.e., linear regression, logit regression, etc.). Increment in probability for a unit increment in a given predictor depends both on the values of the other predictors and the initial value of the given predictors. Therefore, the results can be changed due to the different values of the predictors.

Of note, in the advanced model building phase, MLwiN crashed many times, therefore some of the terms had to be excluded from the initial model. This might have happened due to small number of data points and non-convergence of models.

Conclusion

In Multivariate multilevel model building process there is no satisfactory goodness of fit test yet developed. Therefore it is essential to develop a goodness of fit test in order to access the model adequacy of the multivariate multilevel models. The mortality rates of Asia and Oceania should be reduced, by improving health policies to meet standards like those in North America and Europe. Furthermore, higher consideration should be given to environmental risk factors such as water quality and sanitation to improve personal health.

References

Briggs, D. (2003). Environmental pollution and the global burden of disease. *British Medical Bulletin, 68*(1), 1-24. doi: 10.1093/bmb/ldg019

Carpenter, J. R., Goldstein, H., & Kenward, M. G. (2011). REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software, 45*(5). doi: 10.18637/jss.v045.i05

Climate impacts on Human Health. (n.d.). Retrieved from http://www.epa.gov/climatechange/impacts-adaptation/health.html, Accessed 4 December 2014.

Collett, D. (1991). *Modelling binary data* (1st ed.). London: Chapman & Hall. doi: 10.1007/978-1-4899-4475-7

De Silva, D. & Sooriyarachchi, M. (2012). Generalized Cochran Mantel Haenszel test for multilevel correlated categorical data: an algorithm and R

function. *Journal of the National Science Foundation of Sri Lanka, 40*(2), 137-148. doi: 10.4038/jnsfsr.v40i2.4441

Fodor, J. G., Pfeiffer, C. J. & Papezik, V. S. (1973). Relationship of drinking water quality (hardness-softness) to cardiovascular mortality in Newfoundland. *Canadian Medical Association Journal 108*(11), 1369–1373.

Grilli, L. & Rampichini, C. (2003). Alternative specifications of multivariate multilevel probit ordinal response models. *Journal of Educational and Behavioral Statistics 28*(1), 31-44. doi: 10.3102/10769986028001031

Howell, D. C. (2012). Treatment of Missing Data-Part 1. Retrieved from http://www.uvm.edu/~dhowell/StatPages/Missing_Data/Missing.html, Accessed 5 November 2014.

Jayawardana, N. I. & Sooriyarachchi, M. R. (2014). A multilevel Bayesian analysis of university entrance eligibility for selected districts in Sri Lanka: methods and application to educational data. *Journal of the National Science Foundation of Sri Lanka, 42*(1), 23-36. doi: http://dx.doi.org/10.4038/jnsfsr.v42i1.6676.

Kenny, T. (n.d.). Chronic obstructive pulmonary disease. Retrieved from http://www.patient.co.uk/health/chronic-obstructive-pulmonary-disease-leaflet, Accessed 4 December 2014.

Rasbash, J., Steele, F., Browne, W. J. & Goldstein, H. (2009). *A user's guide to MLwiN, v2.10*. Bristol, UK: University of Bristol.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581-592. doi: 10.1093/biomet/63.3.581

Schols, A. M., Slangen, J., Volovic,s L. & Wouters, E. F. (1998). Weight loss is a reversible factor in the prognosis of chronic obstructive pulmonary disease. *American Journal of Respiratory and Critical Care Medicine, 157*(6), 1791-1797. doi: 10.1164/ajrccm.157.6.9705017

Snijders, T. A. B. & Bosker, R. J. (2012). Multilevel analysis: An introduction to basic and advanced multilevel modeling (2nd ed.). London, UK: Sage Publications Ltd.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ, 338*, b2393. doi: 10.1136/bmj.b2393

What Are the Health Risks of Overweight and Obesity? (13 July, 2012). Retrieved from http://www.nhlbi.nih.gov/health/health-topics/topics/obe/risks, Accessed 2 December 2014.

World Health Organization. (2005). *Preventing chronic diseases: a vital investment.* Retrieved from http://www.who.int/chp/chronic_disease_report/full_report.pdf, Accessed 1 December 2014.

World Health Organization. (2013). *World health statistics 2013*. Retrieved from http://www.who.int/gho/publications/world_health_statistics/2013/en/, Accessed 1 December 2014.

World Life Expectancy. (n.d.) Retrieved from http://www.worldlifeexpectancy.com, Accessed 2 December 2014.

Zhang, J. & Boos, D. D. (1997). Mantel-Haenszel test statistics for correlated binary data. *Biometrics, 53*(4), 1185-1198. doi: 10.2307/2533489