Nepal Journal of Science and Technology 10 (2009) 205-211

Categorical Regression Models with Optimal Scaling for Predicting Indoor Air Pollution Concentrations inside Kitchens in Nepalese Households

Srijan Lal Shrestha
Central Department of Statistics, Tribhuvan University, Kirtipur, Kathmandu
e-mail:srijan_shrestha@yahoo.com

Abstract

Indoor air pollution from biomass fuels is considered as a potential environmental risk factor in developing countries of the world. Exposure to these fuels have been associated to many respiratory and other ailments such as acute lower respiratory infection, chronic obstructive pulmonary disease, asthma, lung cancer, cataract, adverse pregnancy outcomes, etc. The use of biomass fuels is found to be nearly zero in the developed countries but widespread in the developing countries including Nepal. Women and children are the most vulnerable group since they spend a lot of time inside smoky kitchens with biomass fuel burning, inefficient stove and poor ventilation particularly in rural households of Nepal. Measurements of indoor air pollution through monitoring equipment such as high volume sampler, laser dust monitor, etc are expensive, thus not affordable and practicable to use them frequently. In this context, it becomes imperative to use statistical models instead for predicting air pollution concentrations in household kitchens. The present paper has attempted to contribute in this regard by developing some statistical models specifically categorical regression models with optimal scaling for predicting indoor particulate air pollution and carbon monoxide concentrations based upon a cross-sectional survey data of Nepalese households. The common factors found significant for prediction are fuel type, ventilation situation and house types. The highest estimated levels are found to be for those using solid biomass fuels with poor ventilation and Kachhi houses. The estimated PM_{10} and CO levels are found to be 3024 $\mu g/m^3$ and 24115 $\mu g/m^3$ inside kitchen at cooking time which are 5.2 and 40.40 times higher than the lowest predicted values for those using LPG / biogas and living in Pakki houses with improved ventilation, respectively .

Key words: biomass fuel, categorical regression, indoor air pollution, optimal scaling, respiratory ailments

Introduction

Indoor air pollution in the developing world is found to be largely attributed to the use of biomass fuels such as animal dung, crop residues, and wood or coal. Worldwide, approximately 50% of the households and 90% of the rural households use solid fuels for cooking and heating homes (WHO 2002). These fuels are commonly burned in inefficient traditional stoves such as Ageno Chulo in Nepal and inside poorly ventilated kitchens. Consequently, solid fuel use generates substantial emissions of health damaging pollutants such as particulate matter, carbon monoxide, nitrogen oxides, sulfur dioxide, formaldehyde (HCHO), benzene, etc. The disease burden from solid fuel use is most significant in populations with inadequate access to clean fuels such as kerosene, LPG and biogas particularly poor households in rural areas of the developing world. The use of unprocessed fuel is estimated to be nearly zero in developed countries and more than 80% in the countries like China, India, and sub-Saharan Africa (Smith 2002). Conservative estimates of global mortality due to indoor air pollution from solid fuels show that in the year 2000, between 1.5 million and 2 million deaths were attributed to exposure to indoor air pollution This accounts for approximately 4% to 5% of total mortality worldwide (Smith & Mehta 2003). In Nepal, according to population census 2001, nearly

80% households use solid fuels for cooking and heating homes (CBS 2001). Women and children are the most exposed group of individuals because women usually tend to put their babies with them during cooking time.

In households using solid fuels, over a 24h period, the typical mean PM_{10} concentrations can exceed 1000 g/m^3 and carbon monoxide concentrations can exceed 20 ppm (Bruce *et al*. 2000). A cross-sectional study conducted by Nepal Health Research Council (NHRC) in 2003/2004 on health effects from indoor pollution in Nepalese households has found that PM_{10} concentration in kitchens using unprocessed solid fuels (2418 mg/m^3) was about 3 times higher than those using cleaner fuels (792.5 mg/m^3). Similarly, carbon monoxide concentration was 6.67 times higher for solid fuels though mean values were within WHO guidelines of 100,000 mg/m^3 for 15 min averaging time (NHRC 2004). The health effects that have been associated with indoor air pollution are acute lower respiratory infections (ALRI), chronic bronchitis, COPD, asthma, lung cancer, adverse pregnancy outcomes, pulmonary tuberculosis, etc. The analysis of the study data found significant associations between respiratory health outcomes and exposure to biomass smoke. It showed significant odds ratios for COPD / Asthma (3.85) and respiratory symptoms as defined in the British Medical Research Council (BMRC) questionnaire, namely cough (3.71), phlegm (3.08), breathlessness (3.71) and wheezing (5.39). Moreover, the adjusted odds ratios were also found to be significant for cough (3.37), breathlessness (3.47), and wheezing (4.75) when adjusted for smoking. Similarly, the adjusted odds ratios for cough (3.75), phlegm (3.31), breathlessness (3.66), wheezing (5.95), and COPD/asthma (4.18) were also statistically significant when adjusted for age (Shrestha & Shrestha 2005).

Monitoring of air pollutants in every study may not be practicable frequently since the cost required will be naturally much higher. Studies with limited financial resources will have to rely on factors that influence the pollutant concentrations such as fuel type, stove type, ventilation situation, house type, etc. most of which are essentially categorical in nature. As a result, it could be beneficial if we can identify some statistically significant factors and build statistical models for associating these factors with pollutant concentrations. In the paper, attempt has been made in this direction by constructing a couple of statistical models based upon categorical regression with optimal scaling technique (CATREG).

Materials and Methods

A typical statistical model is proposed for predicting indoor air pollutant concentration inside kitchen at the time when fuel is burning. The model is the categorical regression model with optimal scaling technique (CATREG). The model is preferred instead of general linear models (GLM) for the reasons discussed ahead. Even though general linear models are simple models and relatively easy to understand, interpret and apply when the predictor variables are categorical and the response variable is numeric, they rely heavily on several assumptions such as normality, homogeneity in variance across factor level combinations, absence of autocorrelation, etc. Moreover, large sample sizes are required for such models across each of the factor level combinations which may not be fulfilled in practice owing to various constraints mainly financial. Comparatively, even though categorical regression models are relatively complicated and sophisticated involving advanced statistical techniques such as optimal scaling techniques for multivariate categorical data analysis, there are several advantages in using this model as well. One advantage is categorical regression can be run with least assumptions. For instance, normality assumption of the predictor variables is relaxed. In categorical regression since factor levels are coded simultaneously into values, sample sizes need not necessarily be large as in GLM for each factor level combination. In GLM many regression coefficients will be estimated depending upon the number of levels in the categorical factors, p-1 coefficient will be estimated if there are p levels in a predictor variable. In categorical regression, only one coefficient is needed for a predictor variable. Moreover, nonlinear associations can be detected with these models whereas GLMs are basically used for detecting linear associations.

Model specification

In many researches on social, behavioral sciences, marketing, environment and health, etc, variables are often in nominal and ordinal scales. The zero point of the scales used to measure the values is uncertain, the relationships among the different categories is often unknown, and although frequently it can be assumed that the categories are ordered, their mutual distances might still be unknown. The uncertainty in the unit of measurement is not just a matter of measurement error,

because its variability may have a systematic component. An important development in multidimensional data analysis is the optimal assignment of quantitative values to such qualitative scales. This form of optimal quantification (scaling, scoring) is a general approach to treat multivariate categorical data such as in the case of categorical regression and categorical principal component analysis.

Categorical regression quantifies categorical data by assigning numerical values to the categories using optimal scaling method and resulting in an optimal linear regression equation for the transformed variables. Standard linear regression analysis involves minimizing of the sum of squared differences between a response (dependent) variable and a weighted combination of predictor (independent) variables. Variables are typically quantitative, with nominal data recoded to binary or contrast variables. As a result, categorical variables serve to separate groups of cases, and the technique estimates separate sets of parameters for each group. An alternative approach involves regressing the response on the categorical predictor values themselves. Consequently, one coefficient is estimated for each variable. However, for categorical variables, the category values are arbitrary.

Categorical regression extends the standard approach by simultaneously scaling nominal, ordinal, and numerical variables. The procedure quantifies categorical variables so that the quantifications reflect characteristics of the original categories. The procedure treats quantified categorical variables in the same way as numerical variables. Using nonlinear transformations allow variables to be analyzed at a variety of levels to find the best-fitting model.

In the simple linear regression model we wish to predict a response variable **z** from *m* predictor variables in **X**. This objective is achieved by finding a particular linear combination **Xb** that correlates maximally with **z**. This implies minimizing the sum of squared differences between a response (dependent) variable and a weighted combination of predictor (independent) variables. The minimization of the error sum of squares is

$$\| Xb - z \|^2 \quad \text{where} \| Xb - z \| = \sqrt{(Xb - z)'(Xb - z)} \qquad (1)$$

In effect, this maximizes the correlation between the dependent variable **z** and the linear combination of the predictor variables, $\sum_{j=1}^{m} \mathbf{b}_j \mathbf{X}_j$

Incorporating optimal scaling to response as well as predictor variables amounts to the minimization of the expression

$$\| X^*b - z^* \|^2 \quad \text{where} \| X^*b - z^* \| = \sqrt{(X^*b - z^*)'(X^*b - z^*)} \qquad (2)$$

over regression weights **b**, and nonlinear functions $Z^* = \theta(Z)$ and ,j = 1,..., m.

Thus, optimal scaling maximizes the correlation between $\theta(Z)$ and $\sum_{j=1}^{m} b_j \phi_j(X_j)$ over feasible nonlinear functions. These functions are called transformations for quantitative variables and scaling, scorings or quantifications for categorical variables. An alternative approach to linear regression model is therefore, regressing the response on the categorical predictor values themselves. Consequently, one coefficient is estimated for each variable. Model for categorical regression can be expressed as a linear regression model for transformed variables, as given below

$$z^* = X^* b + \varepsilon \qquad (3)$$

where **b** is a vector of standardized coefficients, **ε** is the vector of errors, $\mathbf{X}^*$ is the coefficient matrix containing transformed independent variables, and Z^* is the vector of observations for the transformed response variable. Different optimal scaling levels can be set to the dependent as well as independent variables namely, nominal, spline nominal (transformation is a smooth, possibly non-monotonic, piecewise polynomial of the chosen degree.), ordinal, spline ordinal (transformation is a smooth monotonic piecewise polynomial of the chosen degree) and numeric (Gifi 1990). The details of how categorical variables are dealt with in a framework and how an objective function is optimized in CATREG can be viewed in SPSS white paper (Meulman 1997, 1998) Also many advances on the principles of optimal scaling took place during 1980s. Since the middle 1980s, optimal scaling methods have been extended into more general framework and gradually appeared in mainstream statistical literature (Brieman & Friedman 1985, Ramsay 1989, Buja 1990).

Data

Data for analysis were taken from a cross-sectional household survey conducted under the Nepal Health Research Council project entitled 'Situation analysis of indoor air pollution and development of guidelines for indoor air quality assessment and house building for health' and supported by World Health Organization (WHO), Nepal carried out at some rural and urban areas of Nepal with special focus on rural women who cooked inside unventilated or poorly ventilated kitchens with solid bio-fuels. The survey included 11 Village Development Committees (VDCs) and 4 Municipalities selected randomly from 5 districts of which three were from hills and two from Terai region. From the selected rural and urban regions, 98 households were selected at random with 168 respondents mainly women (94%) who cooked for daily meals. The study was supported by direct measurements of air pollutants mainly particulate matter of size less than 10 micron (PM_{10}) and carbon monoxide (CO). Measurements on other pollutants such as sulfur dioxide (SO_2), nitrogen dioxide (NO_2) and formaldehyde (HCHO) were done on campaign basis in only a fraction of households surveyed. The health responses were judged by an occupational health expert with general health check up including chest examinations, peak flow meter examination and responses obtained through British Medical Research Council (BMRC) questionnaire for respiratory disease identification. The field survey was conducted between November 2003 and February 2004 during dry season in winter. Altitudes of hilly regions included in the survey were about 1,400 meters, and flatlands ranged in altitude from 90 to 240 m (NHRC 2004). Data were analyzed using Statistical Package for the Social Sciences (SPSS) for windows version 13.0.

Model Adequacy Tests

Several measures of model adequacy tests are employed such as goodness of fit by R^2, residual analysis including normality test, homogeneity of variance, residual plots on autocorrelations and partial autocorrelations.

Results and Discussion

The three categorical predictor variables found significant in the model for predicting PM_{10} level and CO level were fuel type, house type and ventilation condition inside kitchen. The description of how these variables were categorized is given in Table 1.

Table 1.Predictor variable description

Variable 1		Variable 2			Variable 3		
Fuel type	Code	House type	Code	Ventilation	Code	Open area(m^2)	Volume(m^3)
Biomass	1	Kachhi	1	Poor	1	0.05 – 1.80	7.0 – 26.0
Kerosene	2	Pakki	2	Moderate	2	0.05 – 1.80	26.0 – 110.0
						1.80– 6.50	**7.0 – 26.0**
LPG/Biogas	3			Improved	3	1.80– 6.50	26 – 110.0

Two models were developed for estimating PM_{10} levels and CO levels separately. The estimated models are:

For predicting PM_{10} concentration

Model 1: $\hat{z}^* = \hat{b}X^* = -0.807\ X^*$

For predicting CO concentration

Model 2: $\hat{z}^* = \hat{b}X^* = -0.819\ X^*$

where functions $Z^* = \theta(Z) = -2.0 + 1.329Z$ for model 1 and $Z^* = \theta(Z) = -5.603 + 0.615Z$ for model 2 are linear functions of the response variables in the models, X* = j (X) is a nonlinear function of the multiple of the categorical variables which is j (Fuel * Ventilation * House type). Total cases used are 79 for model 1 and 77 for model 2. The variables in the fitted models are standardized transformations so that their means are zero and variances are one. The observed values of the response variables in the models are first of all transformed through Box-Cox transformation before modeling. The transformations are used for modeling rather than the original variables in order to stabilize the error variances. The transformations are presented below.

For Model 1:

$$Z = \frac{y^{\lambda} - 1}{\lambda} + k \qquad \text{for } \lambda = 0.01 \text{ and } k = 1.04$$

where y is the observed value of PM_{10} level.

For Model 2: $Z = \frac{y^{\lambda} - 1}{\lambda}$ for $\lambda = 0.01$

where y is the observed value of CO level.

The estimated standardized regression coefficients imply that increase in one standard deviation of the predictor variable (quantified product of fuel, ventilation and house) results in decrease of 0.807 and 0.819 standard deviations of the quantified response variables, respectively. The category quantifications of the response variable (Z^*) and the predictor variable (X^*) are provided in the tables below (Tables 2-5). The estimated standardized regression coefficients with standard errors 0.068 and 0.067 are found to be statistically significant with p=0.000 for model 1 and model 2, respectively. The predicted PM_{10} and CO values for different levels of predictor factor levels are shown in tables 6 and 7.

The highest predicted value for biomass fuel, poor ventilation and kachi house is found to be 5.2 times higher than the lowest predicted value for LPG/biogas fuel, improved ventilation and pakki house type in model 1 for predicting PM_{10} levels. The last column of table 6 shows how the predicted PM_{10} lowers as the levels of the predictor factors changes from biomass fuel to LPG / biogas, poor ventilation to improved ventilation and kachi house type to pakki house type. Similarly, the highest predicted value for biomass fuel, poor ventilation and kachi house is found to be 40.4 times higher than the lowest predicted value for LPG/biogas fuel, improved ventilation and pakki house type in model 2 for predicting CO levels. The last column of table 7 shows how the predicted PM_{10} lowers as the levels of the predictor factors changes from biomass fuel to LPG / biogas, poor ventilation to improved ventilation and kachi house type to pakki house type.

Since the response variables are scaled as numeric, linear associations between the category quantifications and the category arithmetic means are observed for the both models. However, the same is not true for the predictor variables since they are scaled as spline ordinal with 2 degrees of freedom and two interior knots for the models. Consequently, if we examine the graphs between category codes and the corresponding quantifications, we observe nonlinear curves.

Table 2. Category quantification of response variable in model 1

Category	Frequency	Quantification (Z^*)
.01 - .17	2	-1.883
.39 - .77	16	-1.261
.87 - 1.23	17	-.638
1.30 - 1.57	9	-.016
1.72 - 2.15	15	.607
2.20 - 2.68	18	1.229
2.73 - 3.15	2	1.851

Table 3. Category quantification of the predictor variable in model 1

Category	Frequency	Quantification (X^*)
1.00	17	-1.059
2.00	27	-.632
3.00	5	-.161
4.00	5	.323
6.00	7	1.021
8.00	1	1.339
9.00	1	1.360
12.00	8	1.398
18.00	8	1.655

Table 4. Category quantification for response variable in model 2

Category	Frequency	Quantification
5.80	9	-1.888
6.55 - 7.29	5	-1.282
7.73 - 8.48	11	-.676
8.73 - 9.55	16	-.071
9.79 - 10.31	15	.535
10.56 - 11.11	21	1.140

Table 5. Category quantification for predictor variable in model 2

Category	Frequency	Quantification
1.00	17	-1.100
2.00	26	-.521
3.00	6	-.117
4.00	4	.147
6.00	6	.600
8.00	1	.977
9.00	1	1.137
12.00	8	1.515
18.00	8	1.817

Table 6. Predicted PM_{10} levels

Fuel	Ventilation	House	Predicted PM_{10} ($\mu g/m^3$)	% Decrease
Solid Biomass Fuel	Insufficient	Kachi	3024	0.0
Kerosene	Insufficient	Kachi	2337	22.7
Solid Biomass Fuel	Moderate	Kachi	2337	22.7
Solid Biomass Fuel	Insufficient	Pakki	2337	22.7
Solid Biomass Fuel	Improved	Kachi	1758	41.9
Kerosene	Moderate	Kachi	1311	56.6
Kerosene	Insufficient	Pakki	1311	56.6
Solid Biomass Fuel	Moderate	Pakki	1311	56.6
LPG / Biogas	Moderate	Kachi	857	71.7
Kerosene	Improved	Kachi	857	71.7
Solid Biomass Fuel	Improved	Pakki	857	71.7
Kerosene	Moderate	Pakki	706	76.7
LPG / Biogas	Improved	Kachi	697	77.0
LPG / Biogas	Moderate	Pakki	681	77.5
Kerosene	Improved	Pakki	681	77.5
LPG / Biogas	Improved	Pakki	582	80.8

Table 7. Predicted CO levels

Fuel	Ventilation	House Type	Predicted CO ($\mu g/m^3$)	% Decrease
Biomass Fuel	Insufficient	Kachi	24115	0.0
Biomass Fuel	Moderate	Kachi	11698	51.5
Biomass Fuel	Insufficient	Pakki	11698	51.5
Biomass Fuel	Improved	Kachi	7040	70.8
Kerosene	Moderate	Kachi	5046	79.1
Kerosene	Insufficient	Pakki	5046	79.1
Biomass Fuel	Moderate	Pakki	5046	79.1
Kerosene	Improved	Kachi	2841	88.2
LPG/biogas	Moderate	Kachi	2841	88.2
Biomass Fuel	Improved	Pakki	2841	88.2
Kerosene	Moderate	Pakki	1757	92.7
LPG/biogas	Improved	Kachi	1431	94.1
Kerosene	Improved	Pakki	881	96.3
LPG/biogas	Moderate	Pakki	881	96.3
LPG/biogas	Improved	Pakki	597	97.5

Results on model adequacy tests show that the values of adjusted R^2 are found to be 0.637 and 0.658 for model 1 and model 2, respectively. The values may be regarded as moderate. The residuals are found to be normally distributed using Kolmorov-Smirnov nonparametric test ($p=0.208$ for model 1and $p=0.022$ for model 2). The normal probability plots also do not show much deviation from normality for error variables. Studentized residuals are examined with respect to standardized predicted values and two potential outliers were detected (outside ± 2.58) for both the models. The models were rerun after deleting the outliers and the resulting models did not show any outliers. The graphs also show fairly homogenous variance of the residual.

Application of categorical regression models for predicting kitchen PM_{10} concentration and CO concentration during cooking time are found to be suitable when applied to data collected from Nepalese households. The risk factors found statistical significant for predicting the response variables are fuel type, ventilation and house type. The estimated models

predict highest concentrations in kitchens using biomass fuels such as dung, crop residue and wood with poor ventilation and kachhi houses (without concrete use in construction) for the both models. Similarly, the fitted models predict lowest concentrations for both PM_{10} and CO with cleaner fuels (LPG/Biogas) with improved ventilation situation and for pakki those houses. If we closely examine tables 6 and 7 it is encouraging to know that even though biomass fuels are used in kachhi houses which are characteristics of low income people, PM_{10} level and CO level are significantly reduced (41.9% for PM_{10} and 70.8% for CO) if ventilation is improved from poor to improved situation. This is an important research output since switching fuels to LPG or biogas and constructing pakki houses for low and underprivileged section of the people in Nepalese society is difficult. But, if they try to minimize indoor air pollution it can be significantly reduced by improving ventilation condition in the kitchens alone. This can be done by making windows / doors / kitchen volume with larger dimensions with much less economical burden than the economical burden associated with the other two factors. People should be well educated with such findings and also with health consequences of indoor air pollution through awareness programs.

Acknowledgement

The author expresses gratitude to Nepal Health Research Council (NHRC), Kathmandu, Nepal for initiating the project entitled 'Situation analysis of indoor air pollution and development of guidelines for air quality monitoring and house building for health' and World Health Organization (WHO / Nepal) for providing fund and support for the project. Deep appreciation and special thanks go to Dr. Mrigendra Lal Singh, Professor, Dr. Ganga Shrestha, Professor, Dr. Devendra B. Chettri, Professor, and Dr. Iswori Lal Shrestha, Environmental Expert for their encouragements and suggestions. Sincere thanks also go to the study team members, Mr. Sunil Babu Khatri, Environmental Chemist, Mr. Salil devkota, Environmental Engineer and Dr. Sunil Kumar Joshi, Occupational Health Expert for their respective part of contributions in the study.

References

Brieman, L. and J.H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of American Statistical Association* **80**:580-598.

Bruce, N., R. Perez-Padilla and R. Allbak. 2000. *Bulletin WHO 78*, 1078–1092. WHO, Geneva.

Buja, A. 1990. Remarks on functional canonical variates, alternating least squares methods and ACE. *Annals of Statistics* **18**:1032-1069.

CBS 2001. Population census 2001. *Central Bureau of Statistics*, CBS in collaboration with UNFPA, Nepal.

Gifi, A. 1990. *Nonlinear multivariate analysis*. John Wiley and Sons, Chichester.

Meulman, J. 1997. *Optimal scaling methods for multivariate categorical data analysis, SPSS white paper.* Data Theory Group of Social and Behavioral Sciences, Leiden University.

Meulman, J. 1998. *Optimal scaling methods for graphical multivariate data analysis*. Symposium on Computational Statistics, Bristol.

NHRC 2004. Situation analysis of indoor air pollution and development of guidelines for indoor air quality assessment and house building for health. Nepal Health Research Council, Kathmandu.

Ramsay, J.O. 1989. Monotone regression splines in action. *Statistical Science* **4**:425-441.

Shrestha, I.L. and S.L. Shrestha. 2005. Indoor air pollution from biomass fuels and respiratory health of the exposed population in Nepalese households. *International Journal of Occupational and Environmental Health* **11**(2): 150-160.

Smith, K.R. 2002. Indoor air pollution in developing countries: recommendations for research. *Indoor Air* **12**:198-207.

Smith, K.R. and S. Mehta. 2003. The global burden of disease from indoor air pollution in developing countries: Comparison of Estimates. *International Journal of Hygiene and Environmental Health* ***206****:279–289.*

WHO 2002. The health effects of indoor air pollution exposure in developing countries. World Health Organization. Protection of the Human Environment, Geneva.