Lung Cancer Risk Prediction with Stochastic Covariates

Danos D.[1], Oral E.[1], Simonsen N.[2], Fontham E.[2]
[1]LSUHSC School of Public Health, Biostatistics Program,
2020 Gravier Street, New Orleans, LA 70112
[2]LSUHSC School of Public Health, Epidemiology Program,
2020 Gravier Street, New Orleans, LA 70112

Abstract

Lung cancer is the leading cause of cancer death in the US. Previous studies on the nutritional etiology of lung cancer may be inconsistent partially due to inadequate control for smoking through the methods used to do so. Ignoring the stochastic nature of risk factors may contribute to this inconsistency as well. We propose an enhancement of logistic regression analysis that could be used to assess the association of nutritional risk factors with lung cancer. We consider stochastic non-normal covariates and utilize modified maximum likelihood methodology. We show that the proposed estimators are highly efficient, and treating the risk factor as non-stochastic results in loss of efficiency. We illustrate the method using data collected from a population-based case-control study, namely the Lower Mississippi River Interagency Cancer Study (LMRICS), wherein 892 subjects with complete information on diet and smoking habits were interviewed from 1998 to 2001.

Key Words: Modified Maximum Likelihood, Order Statistics, Stochastic Covariates, Logistic Regression, Robustness, Outliers

1. Introduction

In regression models, the covariates have traditionally been assumed to be non-stochastic in nature. In recent years, however, there has been a realization that stochastic covariates are more realistic in practice. Sazak et al. (2006) considered the simple linear regression model

$$E(Y \mid X = x) = \mu_2 + \rho(\sigma_2/\sigma_1)(x - \mu_1), \tag{1.1}$$

where X and Y are stochastic variables. Realize that if both X, and the error term $e = Y - \theta_0 - \theta_1 X_1$ ($\theta_0 = \mu_2 - \mu_1$, $\theta_1 = \rho\sigma_2/\sigma_1$) are distributed from Normal, then the joint distribution of (X, Y) becomes Bivariate Normal distribution. Assuming that both X and the error term are from the Generalized Logistic (GL) distribution

$$GL(b,u): h(u) = \frac{b}{\sigma} \frac{\exp\{-(u-\mu)/\sigma\}}{\left[1 + \exp\{-(u-\mu)/\sigma\}\right]^{b+1}}, \; -\infty < u < \infty, \tag{1.2}$$

where b is the shape parameter, Sazak et al. (2006) derived the modified maximum likelihood estimators (MMLEs) and showed that the MMLEs are more efficient than their

corresponding Least Squares Estimators (LSEs). Likewise, Oral (2004, 2006) considered the binary regression model

$$\pi(x_i) = E(Y_i | X_i = x_i) = F(v_i) \tag{1.3}$$

where X is a risk factor which is stochastic in nature, $v_i = \theta_0 + \theta_1 z_i$, $z_i = (x_i - \mu)/\sigma$ ($\theta_1 \geq 0$), and $F(.)$ is a known cumulative distribution function. Note that the particular choice of $F(.)$ is often arbitrary; however, logistic distribution is commonly used because of the ease of interpretation of the parameter estimates in terms of odds ratios. Oral (2004, 2006) specifically considered the cases where the risk factor is either from the GL family (1.2) or from the Long-Tailed Symmetric (LTS) family

$$LTS(p,u): h(u) = \frac{1}{\sigma\sqrt{k}\ \mathrm{B}(1/2, p-1/2)} \left\{1 + \frac{1}{k}\frac{u^2}{\sigma^2}\right\}^{-p}, \ -\infty < u < \infty, \tag{1.4}$$

where p is the shape parameter, and derived the MMLEs of the model (1.3). She showed that stochastic covariates provide more efficient MMLEs compared to the MMLEs obtained from non-stochastic covariates, and warned that treating the risk factor as non-stochastic results in loss of efficiency. She also showed that the derived MMLEs from stochastic covariates are highly robust with respect to several types of data anomaly.

This study combines the work of Oral (2006) and Sazak et al. (2006) to generalize the model (1.3) for more than one stochastic risk factors.

2. Binary Stochastic Covariates

Consider the same model given in (1.3), where

$v_i = \gamma_0 + \gamma_1 z_{i1} + \gamma_2 z_{i2}$,

$z_{i1} = (x_{i1} - \mu_1)/\sigma_1$, $z_{i2} = (x_{i2} - \mu_2)/\sigma_2$, $\gamma_1, \gamma_2 \geq 0$ and $F(.)$ is the cumulative distribution function of the logistic distribution. Let the covariates X_1 and X_2 have the bivariate distribution as given below

$$h(x_1, x_2) = h_1(x_1) h_2(x_2 \mid x_1), \tag{2.1}$$

where the marginal distribution of X_1 is the GL distribution (1.2) with parameters (μ_1, σ_1, b_1) and the conditional distribution of X_2 given $X_1 = x_1$ is also the GL distribution (1.2) with parameters $(\mu_{2.1}, \sigma_{2.1}, b_2)$; $\mu_{2.1} = \mu_2 + \rho(\sigma_2/\sigma_1)(x_{i1} - \mu_1)$, $\sigma_{2.1} = \sigma_2\sqrt{(1-\rho^2)}$. The joint distribution of X_1 and X_2 involve five parameters which are not functionally related to one another. For the situation above, the full likelihood can be written as

$$L = L_{X_1} L_{X_2|X_1} L_{Y|X_1, X_2}. \tag{2.2}$$

The maximum likelihood estimators obtained from (2.2) are intractable, thus we obtained the MMLEs which are explicit functions of sample observations and, therefore, easy to compute. We are going to report the methodology, our derivations and properties of the derived estimators separately in an upcoming manuscript.

3. Simulation Results

To evaluate the performance of the derived MMLEs, we simulated the means and variances. We generated X_1 from the GL(μ_1, σ_1, b_1) and the conditional distribution of $X_2 \mid x_1$ from the GL$(\mu_{2.1}, \sigma_{2.1}, b_2)$ as explained above. Given in Figures 1.a and 1.b are the simulated values of the means of $\hat{\gamma}_1$ using the MMLEs from stochastic and non-stochastic cases respectively, for *n*=100 and ρ=0.5. True γ_1 values were chosen as 0, 0.5 and 1. Note that the broken lines represent the 5th and 95th percentiles.

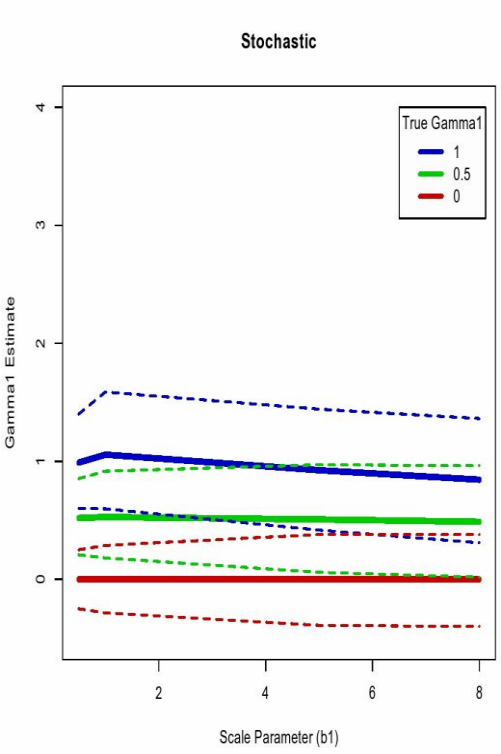

Figure 1.a: Simulated means for $\hat{\gamma}_1$ assuming different b_1 values from stochastic MMLEs.

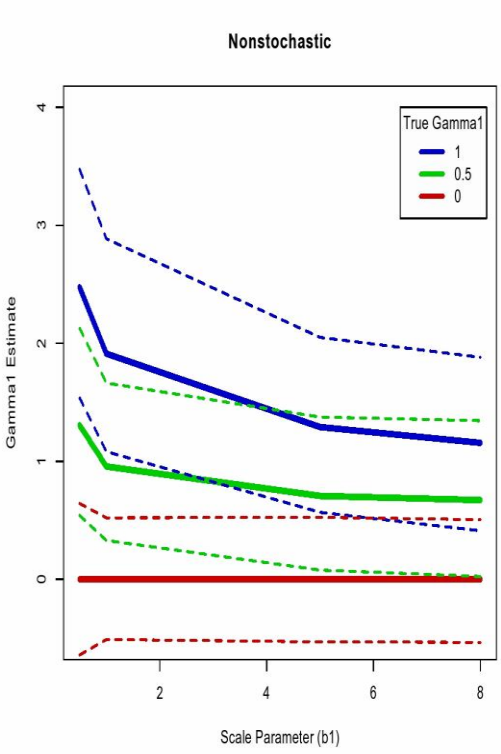

Figure 1.b: Simulated means for $\hat{\gamma}_1$ assuming different b_1 values from non-stochastic MMLEs.

It is clear from Figures 1.a and 1.b that non-stochastic MMLEs give biased estimates. Given in Figures 2.a and 2.b are the simulated values of the means of $\hat{\gamma}_2$ using the MMLEs from stochastic and non-stochastic cases respectively, for *n*=100 and ρ=0.5. True γ_2 values were again chosen as 0, 0.5 and 1. Once more, it is clear from Figures 2.a and 2.b that stochastic MMLEs give better estimates. We also simulated the variances of the derived estimators and compared them with the MMLEs which we obtained from the

non-stochastic case. We assumed that ρ=0.5 and considered several values for (b_1, b_2). We provide our results for *n*=30 and *n*=100 in Table 1 below.

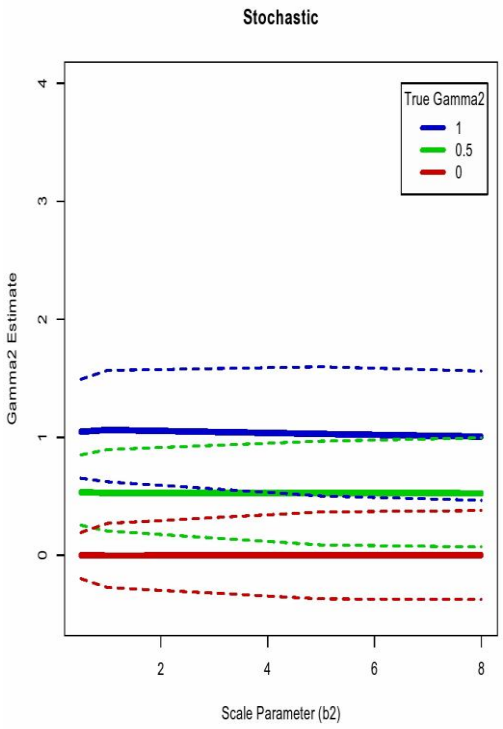

Figure 2.a: Simulated means for $\hat{\gamma}_2$ assuming different b_1 values from stochastic MMLEs.

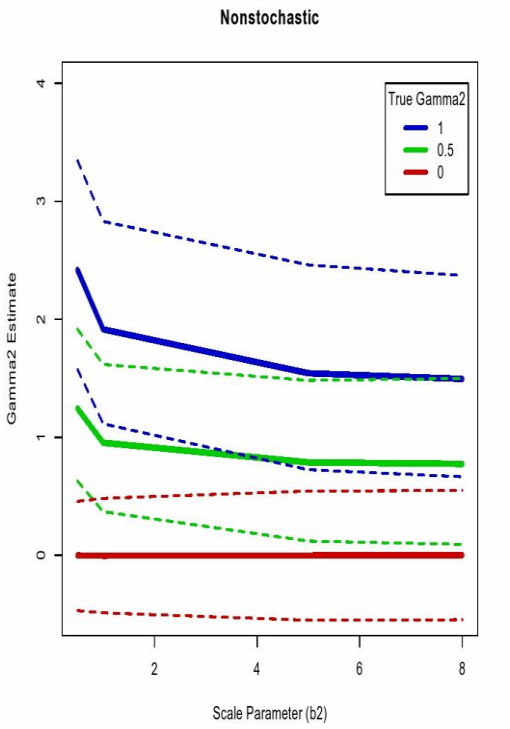

Figure 2.b: Simulated means for $\hat{\gamma}_2$ assuming different b_1 values from non-stochastic MMLEs.

Table 1: Simulated means and variances of the MMLEs under both non-stochastic and stochastic assumptions for different combinations of $b_1, b_2, \gamma_0, \gamma_1, \gamma_2$ and *n*.

(*b*1,*b*2)	**n=30**					
	$\gamma_0 = 0$		$\gamma_1 = 0.5$		$\gamma_2 = 1.0$	
(0.5,0.5)	Mean	Variance	Mean	Variance	Mean	Variance
Stoch	0.050	0.885	0.527	0.117	1.074	0.180
Non-stoch	-2.601	0.804	1.328	0.760	2.674	1.155
	$\gamma_0 = 0$		$\gamma_1 = 1.0$		$\gamma_2 = 0.5$	
(0.5,0.5)	Mean	Variance	Mean	Variance	Mean	Variance
Stoch	0.043	0.862	1.074	0.198	0.551	0.107
Non-stoch	-2.404	0.845	2.712	1.301	1.370	0.663
	$\gamma_0 = 0$		$\gamma_1 = 0.5$		$\gamma_2 = 1.0$	
(5.0,5.0)	Mean	Variance	Mean	Variance	Mean	Variance
Stoch	0.396	3.334	0.462	0.346	0.923	0.456
Non-stoch	3.892	0.826	0.625	0.653	1.241	0.843
	$\gamma_0 = 0$		$\gamma_1 = 1.0$		$\gamma_2 = 0.5$	
(5.0,5.0)	Mean	Variance	Mean	Variance	Mean	Variance
Stoch	0.279	3.786	0.966	0.563	0.487	0.348
Non-stoch	3.609	0.859	1.314	1.060	0.655	0.623

(*b*1,*b*2)	n=100					
	$\gamma_0 = 0$		$\gamma_1 = 0.5$		$\gamma_2 = 1.0$	
(0.5,0.5)	Mean	Variance	Mean	Variance	Mean	Variance
Stoch	0.001	0.222	0.488	0.031	0.981	0.051
Non-stoch	-2.504	0.258	1.247	0.207	2.497	0.343
	$\gamma_0 = 0$		$\gamma_1 = 1.0$		$\gamma_2 = 0.5$	
(0.5,0.5)	Mean	Variance	Mean	Variance	Mean	Variance
Stoch	0.019	0.221	0.994	0.055	0.503	0.027
Non-stoch	-2.289	0.262	2.539	0.378	1.277	0.178
	$\gamma_0 = 0$		$\gamma_1 = 0.5$		$\gamma_2 = 1.0$	
(5.0,5.0)	Mean	Variance	Mean	Variance	Mean	Variance
Stoch	0.161	0.835	0.471	0.146	0.945	0.166
Non-stoch	3.804	0.371	0.643	0.274	1.285	0.310
	$\gamma_0 = 0$		$\gamma_1 = 1.0$		$\gamma_2 = 0.5$	
(5.0,5.0)	Mean	Variance	Mean	Variance	Mean	Variance
Stoch	0.127	0.691	0.942	0.144	0.479	0.110
Non-stoch	3.434	0.336	1.286	0.273	0.651	0.204

From Table 1 we conclude that assuming non-stochasticity for risk factors which are in fact stochastic in nature yields inefficient estimators.

4. Lung Cancer Study (LMRICS)

We applied the methodology to data from the Lower Mississippi River Interagency Cancer Study (LMRICS). LMRICS included a population-based case-control study of lung cancer in the Louisiana industrial corridor encompassing 11 parishes along the Mississippi river. Newly diagnosed incident cases aged 20-74 were enrolled along with an equal sample of corresponding controls frequency matched on race, gender and five year age group using stratified random sampling (Simonsen et al., 2010). The study focused on potential exposure to environmental carcinogens through proximity to petrochemical sites, but collected extensive data on other potential risk factors including smoking history, physical measures, occupation and diet. Interviews conducted from 1998 through 1991 yielded a total of 892 subjects with complete information on diet and smoking habits.

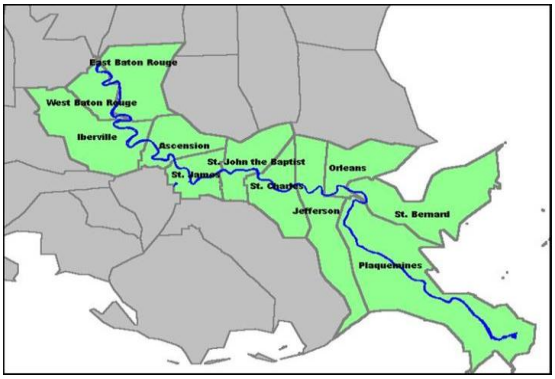

Figure 3: LMRICS study area.

Information on diet was obtained from a food-frequency questionnaire administered to study participants in the course of an interview. The nutrition data from the LMRICS study involved many highly skewed distributions. As an example, we provide the histograms for body mass index (BMI) as well as dietary intake of three nutrients from the study below. From Figure 4, it can be seen that BMI, fat, folates and protein all have positively skewed distributions with outliers on the tails of the distributions.

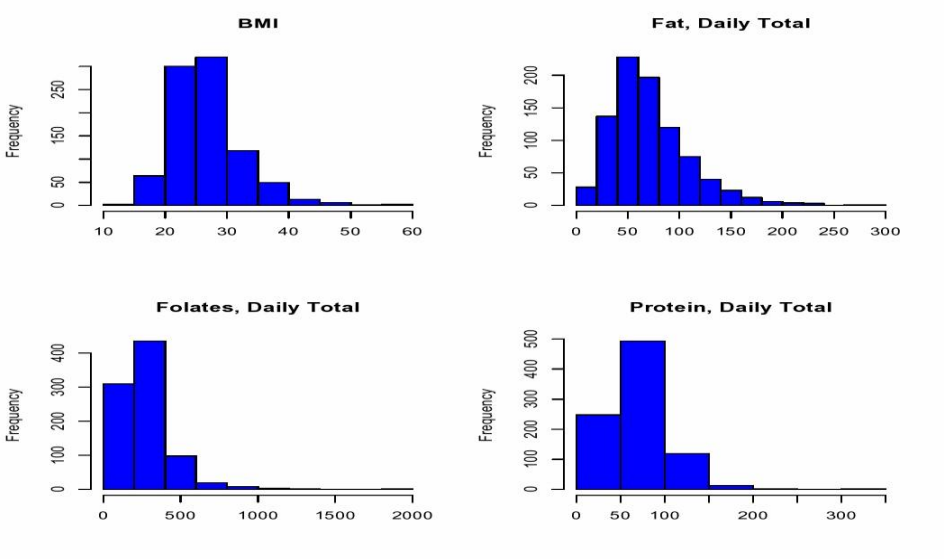

Figure 4: Histograms of BMI and three dietary factors from the LMRICS study

As an illustration of the methodology, we specifically considered BMI and protein as the bivariate risk factors. We modeled the marginal and the conditional distributions of BMI and protein (given BMI) with GL($b_1 = 5$) and GL($b_2 = 8$) respectively. We give the histograms and the corresponding Q-Q plots in Figure 5. We calculated the MMLEs under both non-stochastic and stochastic assumptions. Note that the MMLEs for the non-stochastic case correspond to the traditional maximum likelihood estimators (MLEs).

Since smoking is the main risk factor for lung cancer, any assessment of the association between that cancer and a dietary factor like protein intake would need to take smoking habits into account as well. For illustration purposes we first provide the results of crude logistic regression models including only BMI and protein (Table 2). We then provide results adjusted for smoking through the inclusion of a simple binary categorical variable (Smoked) for having ever been a smoker (Table 3).

Table 2: Logistic regression analyses for the crude model. BMI and protein are modeled with a bivariate non-normal distribution.

	Non-stochastic MLEs			Stochastic MMLEs		
	Estimate	Standard Error*	z	Estimate	Standard Error*	z
Intercept	0.031	0.070	0.443	0.353	0.182	1.940
BMI	-0.437	0.128	-3.414	-0.307	0.062	-4.952
Protein	0.176	0.090	1.956	0.123	0.052	2.365

*The standard errors are obtained via bootstrapping.

From Table 2, it can be seen that the standard errors of the BMI and protein risk factors significantly decrease under the stochastic assumption. As expected, the associated Wald statistics (z) become larger.

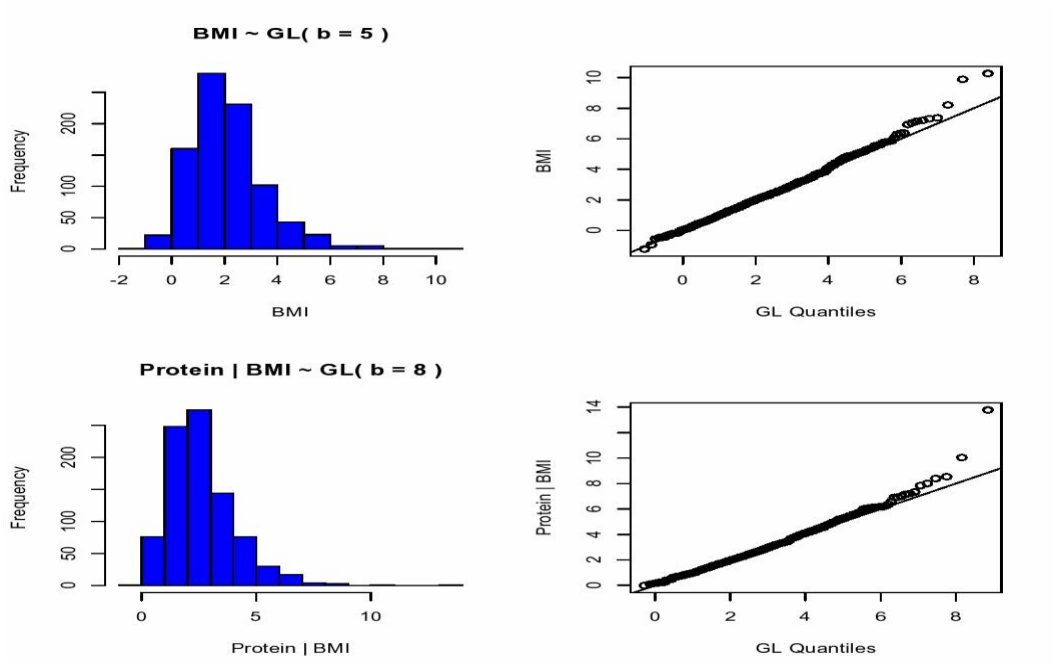

Figure 5: Marginal distribution of BMI and conditional distribution of protein with their corresponding Q-Q Plots.

Table 3: Logistic regression analyses for the adjusted model. BMI and protein are modeled with a bivariate non-normal distribution, Smoked is assumed to be a non-stochastic variable.

	Non-stochastic MLE			Stochastic MMLE		
	Estimate	Standard Error*	z	Estimate	Standard Error*	z
Intercept	-1.718	0.226	-7.605	-1.241	0.258	-4.812
Smoked	2.112	0.241	8.775	2.105	0.235	8.976
BMI	-0.403	0.087	-4.620	-0.282	0.065	-4.331
Protein	0.066	0.077	0.852	0.047	0.054	0.878

*The standard errors are obtained via bootstrapping.

From Table 3, where we assumed that having smoked is a non-stochastic variable, it can be seen that the standard errors of the BMI and protein obtained under the stochastic assumption are still smaller with respect to their corresponding standard errors that are obtained under the non-stochastic assumption. It is interesting to observe that the standard error of the risk factor smoked stays almost the same under both models, which is expected since both models assume that it is a non-stochastic risk factor.

5. Concluding Remarks

Traditionally in the binary regression literature, the risk factors have been assumed to be non-stochastic. This approach, however, is too restrictive for real-life applications. We give solutions for the situations where risk factors are not necessarily non-stochastic in nature. As an example, studies of nutritional factors in lung cancer have produced largely inconclusive results, and typically treated these factors as non-stochastic. Using data on protein intake and BMI from a lung cancer case-control study, the stochastic approach produces substantially different estimates. Treating risk factors as non-stochastic results in loss of efficiency if they are in fact stochastic.

References

Oral, E., and Gunay S. (2004) Stochastic Covariates in Binary Regression, Hacettepe Journal of Mathematics and Statistics, 33: 97-109.

Oral, E. (2006) Binary Regression with Stochastic Covariates, Communications in Statistics-Theory and Methods, 35: 1429-1447.

Sazak, H. S., and Tiku, M. L., and Islam, M. Q. (2006) Regression Analysis with a Stochastic Design Variable. International Statistical Review, 74(1): 77-88.

Simonsen, N., Scribner, R., Su, L.J., Williams, D., Luckett, B.,Yang, T., Fontham, E. T. H. (2010) Environmental Exposure to Emissions from Petrochemical Sites and Lung Cancer: The Lower Mississippi Interagency Cancer Study. Journal of Environmental and Public Health.