Gastroenterology and Hepatology From Bed to Bench.

ORIGINAL ARTICLE

Application of semi-parametric single-index two-part regression and parametric two-part regression in estimation of the cost of functional gastrointestinal disorders

Mohadese shojai[1], Anoshirvan Kazemnejad[1], Farid Zayeri[2], Mohsen Vahedi[3]

[1] *Biostatistics Department, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran*

[2] *Department of Biostatistics, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran*

[3] *Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran.*

ABSTRACT

Aim: For the purpose of cost modeling, the semi-parametric single-index two-part model was utilized in the paper. Furthermore, as functional gastrointestinal diseases which are well-known as common causes of illness among the society people in terms of both the number of patients and prevalence in a specific time interval, this research estimated the average cost of functional gastrointestinal diseases.

Background: Health care policy-makers seek for real and accurate estimations of society's future medical costs. However, data dealt with in hygienic studies have characteristics which make their analysis complicated; distribution of cost data is highly skewed since many patients pay great costs. In addition, medical costs of many persons are zero in a specific time interval. Indeed, medical costs data are often right skewed, including remarkable number of zeros, and may be distributed non-homogeneously.

Patients and methods: In modeling these costs by the semi-parametric single-index two-part model, parameters were determined by method of least squares; a result of this method was compared with the results yielded from two-part parametric model.

Results: Average costs of functional gastrointestinal diseases and their standard deviation in semi-parametric and parametric methods were yielded as \$72.69±108.96 (R^2=0.38) and \$75.93±122.29 (R^2=0.33) respectively.

Conclusion: Based on R^2 index, the semi-parametric model is recognized as the best model. Totally, the two-part parametric regression model is a simple and available model which can be easily interpreted; on the other hand, though the single-index two-part semi-parametric model cannot be easily interpreted, it has considerable flexibility. The study goals can be indeed used as the main factor for choosing one of these two models.

Keywords: Semi-parametric regression, Two-part model, Single-index model, Semi-parametric least squares.

(Please cite as: **Shojai M, Kazemnejad A, Zayeri F, Vahedi M. Application of semi-parametric single-index two-part regression and parametric two-part regression in estimation of the cost of functional gastrointestinal disorders. Gastroenterol Hepatol Bed Bench 2013;6(4):202-209).**

Introduction

Health care policy makers seek realistic deductions and precise estimation of future medical costs of the communities (1). The topic of modifying health cares and interest in the development of national care plans are our motivations in understanding the relative effects of demographic indices, different levels of the diseases, and the change in the type of services provided on the disease costs in practical models (2). Functional gastrointestinal (GI) disorders are

Received: 30 April 2013 *Accepted*: 20 June 2013
Reprint or Correspondence: Anoshirvan Kazemnejad, PhD. Biostatistics Department, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
E-mail: Kazem_an@modares.ac.ir

common in many countries. Thus, it is expected that the disorders have a considerable economic burden for the countries. These disorders lead to higher rate of receiving health cares, and also affect the people's job and their productivity (3). The chronic nature of GI diseases and the uncertain points in their diagnosis and treatment increase the costs (4). According to different studies between 2004 and 2009 carried out in various countries, the prevalence rates of dyspepsia, irritable bowel syndrome, and gastroesophageal reflux are 11-27%, 10-15%, and 18-40%, respectively (5-7).

Furthermore, the data in health studies have characteristics that complicate their analysis. The data distribution is highly skewed, due to the high costs of some patients. Also, the medical costs of some people in a specific time interval are zero. Therefore no simple parametric distribution is suitable to describe such semi-continuous data. When the goal is to determine the average cost of patients' population for a certain disease, without taking these characteristics into account, the estimations and statistical analyses would be inaccurate. For the first time in the health economics and particularly in the studies related to the health costs, multi-part models were introduced by Duan (8). With a high degree of flexibility, the two-part model in fact allows to analyze the zero and positive costs in two separate processes (2, 8-11). In the two-part parametric regression model, the semi-parametric single-index regression is fitted to the positive costs in the second part. This is while in the semi-parametric single-index two-part regression, a semi-parametric single-index regression is fitted to the positive costs in the second part. The semi-parametric single-index model is certainly one of the most important semi-parametric regression models, which can be considered as a generalization of the generalized linear models. This model does not include many limiting assumptions of the parametric family models, for estimation of the conditional mean function. Meanwhile, the model preserves many favorable characteristics of linear models and the least-squares method. The model does not have limitations of the parametric models, and also to some extent compensates the flexibility of nonparametric models (12-14).

High costs of health care for each patient make these diseases a considerable resource for the health care costs, and thus a potential target in reducing the costs. Although the economic costs of GI disorders have been studied in many developed countries, in this respect there is few data available from countries such as Iran.

In the study, using fitness of the two models of parametric two-part and semi-parametric single-index two-part models on the data related to functional GI disorders, we have tried to obtain an accurate estimation of the average costs of the disease and identify the variables affecting the costs. The data obtained from these two models were compared to each other according to the R^2 goodness of fit to present the most appropriate model for estimation of the GI disease costs.

Patients and Methods

Data related to functional GI disorders

During 2006-2007, the data related to 2929 patients with functional GI disorders were collected by convenient sampling in a cross-sectional manner in the Gastrointestinal and Liver Research Center, Shahid Beheshti Medical University, Tehran. The project was approved by the ethics committee of the Gastrointestinal and Liver Research Center, Taleghani Hospital. For all the patients, the data related to the variables of age, sex (male/ female), education level (M.Sc. and higher degrees, B.Sc., high school diploma, below high school diploma, and primary school), marital status (single, divorced, widow, and married), ability to work during the disease episodes (never, to some extent, moderate, high, very high), number of visits by general physician and specialist, number of diagnostic tests, health insurance coverage (not covered, state insurance, complementary insurance), and the patients' costs were collected. The data were used in the analysis.

In the study, the sum of direct and indirect GI disorder costs was considered as the total cost of each patient. Direct costs are the costs an individual pays directly, including the costs of physician visit, drugs, and laboratory and paramedical test. The costs were calculated in dollars. Indirect costs addresses the costs that cannot be calculated directly, and includes the patients' lost income, and the economic value of insensible costs such as death, pain, quality of life, and even inconveniences. In fact, the indirect costs are the lost resulted from the patients' inconveniences, which is calculated by multiplying the average daily income of an employed person by the number of missed work days or days of inconvenience. The average daily income of an employed person was calculated according to the data provided by the World Bank. The costs of each individual were recorded according to economic indices, in international dollars (purchasing power parity dollars (PPP$).

Data analysis

In this paper, we have tried to model the costs of functional GI disorders. In the study, the costs of 21% of the individuals were zero, while costs of almost 1% of the people studied exceeded 1000 PPP$. Therefore no simple parametric distribution is suitable to describe such data. To this end, we have employed parametric two-part regression and semi-parametric single-index two-part regression models. In both models, in the first part, the models were fitted the two-value status costs (zero and positive) of the logistic regression. In the second part, first a multiple-variable linear regression and then a semi-parametric single-index regression were fitted to the positive costs, and the results obtained from the two models were compared. In the first part of the model, we used logistic regression to determine the relationship between the two-valued costs on the one hand, and independent variables of age, sex, education level, marital status, ability to work during the disease episodes, number of visits by general physician and specialist, number of diagnostic tests, and health insurance coverage on the other hand. Also, the probability of having positive costs was calculated for each patient. In the second part of the model, we used semi-parametric single-index and multiple parametric regression to evaluate the relationship between positive costs on the one hand and independent variables of ability to work during the disease episodes, number of missed work days or the days with reduced productivity, and days of hospitalization on the other hand. The average costs expected for each individual was calculated. By multiplying the results obtained for each part of the model, the final estimate of costs was calculated for each patient. The variables affecting the costs were also determined in the first and second parts of the model. The parameters were estimated by the least-squares and semi-parametric least-squares (SLS) methods in the parametric and semi-parametric single-index models, respectively. The results obtained from the two models were compared using the R^2 goodness of fit index. The data were analyzed using SPSS version 16, and the SIM package in the R software. In all steps, level of significance was considered 0.05.

Results

After revising the data, those related to 1907 patients remained in the study. The patients included in the study were in the age range of 20-80 years.

Table1. Results of multiple logistic regression between having positive costs and other variables

Variable	Category	Estimate	SE†	P-Value	OR‡
Age (45.8±14.8*)	-	-0.006	0.005	0.225	0.994
Sex	Female (n=1093, 57.3%)	0.088	0.140	0.527	1.642
	Male (n=814, 42.7%)		reference		
education level	Master's degree or higher (n=35, 1.8%)	0.255	0.587	0.664	1.290
	Bachelor (n=259, 13.6%)	-1.033	0.244	<0.001	0.356
	high school (n=528, 27.7%)	-0.958	0.208	<0.001	0.384
	Less than high school (n=605, 31.7%)	-0.596	0.202	0.003	0.551
	Primary (n=480, 25.2%)		reference		
Marital Status	Single (n=198, 10.4%)	1.927	0.488	<0.001	6.870
	Divorced (n=1557, 81.6%)	1.825	0.451	<0.001	6.203
	Widow (n=122, 6.4%)	1.809	0.537	<0.001	6.109
	Married (n=30, 1.6%)		reference		
ability to work	Never (n=1108, 27.5%)	-0.742	0.635	0.242	0.476
	Low (n=384, 20.1%)	-0.285	0.648	0.660	0.752
	Average (307, 16.1%)	-0.563	0.652	0.383	0.567
	High (n=61, 3.2%)	-0.231	0.745	0.756	0.794
	Very much (n=48, 2.5%)		reference		
number of diseases (3±1.5*)	-	0.100	0.042	0.017	1.105
number of missed work days (1.3±6.7*)	-	0.179	0.046	<0.001	1.196
number of physician visits (1.7 ± 2.2*)	-	0.613	0.061	<0.001	1.845
insurance coverage	No (n=525, 27.5%)	1.920	0.582	<0.001	6.819
	State Insurance (n= 1363, 71.5%)	1.356	0.588	0.021	3.879
	Supplementary insurance (n=19, 1.0%)		reference		

* Mean ± standard error; †Standard error; ‡ Odds ratio

Table 2. Results of multiple linear regression and semi-parametric regression and between the amount of positive costs and other variables

model	Variable	Estimate	SE*	P-Value
parametric two-part	Constant	31.206	6.902	<0.001
	number of missed work days and days with reduced productivity	14.309	2.236	<0.001
	number of visits by general physician and specialist	12.437	0.689	<0.001
	frequency of hospitalization	453.676	24.827	<0.001
semi-parametric two-part	number of missed work days and days with reduced productivity	1	-	-
	number of visits by general physician and specialist	6.299	0.096	<0.001
	frequency of hospitalization	191.707	7.189	<0.001

* Standard error

Table 3. Results of estimation of GI disease costs with different approaches

costs	Mean	Standard error	Min	Max	R^{2*}
real costs	78.35	222.36	0	5183.81	-
cost estimations by the parametric two-part	75.93	122.29	2.89	1394.22	0.33
cost estimations by the semi-parametric two-part	72.69	108.96	5.66	1546.39	0.38

* Coefficient of determination

The average cost for individuals studied was 78.35 ± 222.37 PPP$ (Table 1).

By fitting logistic regression, the variables of education level, number of diseases the patients have, number of missed work days or days with reduced productivity, number of physician visits, and insurance coverage remained in the model ($p<0.05$). However, since we were interested to evaluate the effect of variables of age, sex, and ability to work during the disease episodes, we kept these variables in the model (Table 1). By fitting this model, the probability of having positive costs was separately calculated for each patient to be used in following steps.

In the second part of the model, linear regression and a semi-parametric regression we reused to fit the positive costs. The results obtained from the two models were then compared (Table 2).

By fitting the model, without taking into account that the individual may have zero cost, the average disease cost was determined for each person. In the next step, multiple linear regression and semi-parametric regression was used to fit the positive costs and number of missed work days and days with reduced productivity, number of visits by general physician and specialist, and frequency of hospitalization (Table 2). According to the results obtained in the study, if the conditions remain fixed in the linear regression, for each hospitalization, the costs of patient would be 454 PPP$. Each missed work day would costs 12 PPP$) to the patient. Also, each visit by a general physician or specialist would cost 14 PPP$. The semi-parametric regression coefficients for the variables of frequency of visits by a general physician or specialist, number of missed work days, and hospitalization were determined as 1 (in the least-squares method, the coefficient for the first variable is estimated as 1), 6.299, and 191.707, respectively. In this semi-parametric regression model, similar to the parametric regression, the independent variable coefficient indicates a change in the independent variable corresponding to the change in the dependent variable. However, the exact level of change cannot be specified.

By combining the results of the two steps, the final average cost for each individual was calculated. The cost estimations by the parametric two-part and semi-parametric two-part models are provided in Table 3. The results could be compared with the real costs (Table 3).

Discussion

In the parametric two-part regression model, the average costs of GI disorders were estimated as 75.93 PPP$ with the standard deviation of 122.29. The minimum and maximum costs were determined 2.89 and 1394.32 PPP$, respectively. In the semi-parametric single-index two-part regression model, the mean and standard deviation of GI disorder costs were determined to be 72.69 and 108.96 PPP$. The minimum and maximum costs in the semi-parametric regression model were estimated as 5.66 and 1546.39 PPP$, respectively. The mean true costs in the samples studied were 78.35 PPP$, with the standard error of 222.36 PPP$. The minimum and maximum costs in the samples studied were 0 and 5183.81 PPP$, respectively.

In Iran, health cost is of great importance, as the Iranian population is young and maintenance of the workforce and expansion of general and higher education would be helpful in economic growth and development. Many studies have been carried out on economic burden of different diseases in Iran. However, almost all these studies have employed conventional economic measures to calculate the direct and indirect costs of a certain disease, and in all the studies no statistical method has been used to determine the disease average costs. In previous years, the costs of different diseases including diabetes, diabetic nephropathy, lung cancer, and hip fracture have been evaluated using economic methods. With regard to the costs of the GI diseases, Roshandel et al. reported the direct costs for consulter and non-consulter patients in purchasing power parity dollars (PPP$) as 92.04 and 1.04 for IBS, 100.94 and 0.39 for unspecified functional bowel disorder (FBD), 57.23 and 1.04 for constipation, and 71.35 and 0.63 for abdominal bloating, respectively. Indirect costs (for consulters and non-consulters) were IBS (811.85, 669.09), unspecified FBD (705.85, 263.47), constipation (587.48, 97.49), and abdominal bloating (147.88, 38.60), respectively (15). In another study, Moghimi et al. reported the overall cost of dyspepsia and gastroesophageal reflux (GERD) as 120.2 and 111.4 PPP$, respectively. Furthermore, they reported the direct costs of dyspepsia and GERD per individual per year as 108 and 98 PPP$, respectively (16). In a similar study, using economic methods, Lashkajani et al. reported the cost of dyspepsia and GERD in the range of 172-176 PPP$ (4). While few population-based studies on the economic burden of functional bowel disorders (FBD) have been published from developing countries like Iran, Moghimi et al estimated direct and indirect costs for five groups of patients: irritable bowel syndrome (IBS), functional constipation (FC), unspecified-FBD (U-FBD), functional abdominal bloating (FAB) and functional diarrhea (FD). They concluded that the highest proportion of drug consumption was in IBS patients. The highest mean duration of absence from work was seen in IBS patients (2.26 days). A higher indirect cost of illness was found in IBS (54 PPP$), whereas it was zero in FD. These results showed that the economic burden of FBD seems to be moderately high in Iran and it imposes a relatively heavy financial burden on the Iranian national health system because of its high prevalence and its impact on quality of life, productivity and waste of resources (17). Furthermore direct and indirect costs of functional constipation were calculated during 2006 to 2007 by mohaghegh et al. Of the total 18,180 participants in this study, 435 (2.4%) had FC according to Rome III criteria. The results showed that although the economic burden of FC does not seem to be substantial in comparison to other major health problems, it still exacts a substantial toll on the health system for two reasons: chronicity and ambiguity of symptoms (18). While soruri reported the functional bowel disorders in Iranian population using Rome III criteria by cross-sectional household survey from 2006 to 2007 in Tehran province. In the multivariate analysis, women had a higher risk of any functional bowel disorders FBDs than men, except for functional diarrhea (FD). Their studies revealed a low rate of FBDs among the urban population of Tehran province. The ROME III criteria itself, and the problems with interpretation of the data collection tool may had contributed in underestimating the prevalence of FBD. In addition the reliability of recall over 6 months in Rome III criteria was questionable for their population (19). In the current study, we have tried to provide an estimation of the costs of functional GI disorders using a statistical approach.

As it was mentioned earlier, an outstanding characteristic of the health care costs is a scattered data with a significant proportion of zero values. This factor should be taken into account when modeling these data. With its high flexibility, the two-part model gives the best fit for costs with a large proportion of zero values and highly skewed data, and provides appropriate and feasible estimations. The model has been evaluated in different studies.

It seems that making use of these models overcomes the problems of conventional parametric models, as the data available on the GI disease costs contain a considerable proportion of zero values. Also, the high costs make the use of conventional parametric models impossible.

According to the results obtained in the study, with regard to the true mean and the corresponding value estimated, it was observed that the two values were close to each other, and the standard deviation was reduced in both methods. Furthermore, the mean cost is not equal to zero in the two methods, and the idea of maximization of the zero costs was accomplished. The minimum cost in the parametric regression model was 2.89 PPP$, which was lower than the corresponding value in the semi-parametric model. In the sample studied, approximately1% of the individuals had medical costs exceeding 1000

PPP$. The percentage was reduced in both methods; such that the maximum cost did not reach 2000 PPP$. This is while the value was in fact above 5000 PPP$(5183.81 PPP$).

To compare the models, we employed the R^2 goodness of fit. The R^2 value was 0.33 and 0.38 for the two-part parametric and semi-parametric single-index two-part models, respectively. With regard to this index, the semi-parametric single-index two-part model was the most appropriate model. Considering the results obtained both methods yielded satisfactory values for goodness of fit. In fact, the notable characteristics of the two methods indicate that the researchers should select the model according to their main objectives. The parametric single-index two-part model is a simple and available model, which can be easily interpreted. This is while although the semi-parametric single-index two-part model cannot be interpreted easily, it shows a considerable level of flexibility. In fact, the research objectives can be considered as the main factor for selection of one of these two models.

Acknowledgements

The authors wish to thank the staff of the Gastrointestinal and Liver Research Center, Shahid Beheshti Medical University, Tehran. The paper is derived from the M.Sc. thesis of Biostatistics, and was fulfilled by the financial support of the Tarbiat Modares University, Tehran.

References

1. Zhou X-H, Liang H. Semi-parametric single-index two-part regression models. Comput Stat Data Anal 2006;50:1378-90.

2. Liu L, Strawderman RL, Cowen ME, Shih Y-CT. A flexible two-part random effects model for correlated medical costs. J Health Econ 2010;29:110-23.

3. Ghadir MR, Ghanouni AR. A review of the treatment of irritable bowel syndrome. Journal of Ghom University of Medical Science 2010; 2: 59-66. [In Persian].

4. Rezailashkajani M, Roshandel D, Shafaee S, Zali M. A cost analysis of gastro-oesophageal reflux disease and dyspepsia in Iran. Dig Liver Dis 2008;40:412-17.

5. Nyrop K A, Palsson O S, Levy RL, Korff MV, Feld AD, Turner M J, et al. Costs of health care for irritable bowel syndrome, chronic constipation, functional diarrhea and functional abdominal pain. Aliment Pharmacol Ther 2007; 26: 237–48.

6. Schwenkglenks M, Marbet UA, Szucs TD. Epidemiology and costs of gastroesophageal reflux disease in Switzerland: a population-based study. Soz Praventivmed 2004;49:51-61.

7. Hoveyda N, Heneghan C, Mahtani KR, Perera R, Roberts N, Glasziou P. A systematic review and meta-analysis: probiotics in the treatment of irritable bowel syndrome. BMC Gastroenterol 2009;9:15.

8. Duan N, Manning WG, Morris CN, Newhouse JP. A comparison of alternative models for the demand for medical care. J Bus Eco Stat 1983;1:115-26.

9. Madden D. Sample selection versus two-part models revisited: the case of female smoking and drinking. J Health Econ 2008; 27: 300-307.

10. Mullahy J. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. J Health Econ 1998;17:247-81.

11. Buntin MB, Zaslavsky AM. Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. J Health Econ 2004;23:525-542.

12. Geenens G, Delecroxi M. A survey about single-index models theory. Int J Stat and Syst 2006; 1:203-30.

13. Horowitz J, Horowitz J. Semiparametric methods in econometrics. New York: Springer; 1998.

14. Liu A, Kronmal R, Zhou XH. Semiparametric two part models with proportionality constraints: analysis of the multi-ethnic study of atherosclerosis (MESA). Washington: University of Washington Biostatistics Working Paper Series; 2009.

15. Roshandel D, Rezailashkajani M, Shafaee S, Zali MR. A cost analysis of functional bowel disorders in Iran. Int J Colorectal Dis 2007;22:791-99.

16. Moghimi-Dehkordi B, Vahedi M, Khoshkrood-Mansoori B, Kasaeian A, Safaee A, Habibi M, et al. Economic burden of gastro-oesophageal reflux disease and dyspepsia: A community-based study. Arab J Gastroenterol 2011;12:86-89.

17. Moghimi-Dehkordi B, Vahedi M, Pourhoseingholi MA, Khoshroom-Mansoori B, Safaee A, Habibi M, et al. Economic burden attributable to functional bowel disorders in Iran: A cross-sectional population-based study. J Dig Dis 2011;12:384-92.

18. Sorouri M, Pourhoseingholi MA, Vahedi M, Safaee A, Moghimi-Dehkordi B, Pourhoseingholi A, et al. Functional bowel disorders in Iranian population using Rome III criteria. Saudi J Gastroenterol 2010;16:154-60.

19. Mohaghegh Shalmani H, Soori H, Mansoori BK, Vahedi M, Moghimi-Dehkordi B, Pourhoseingholi MA, et al. Direct and indirect medical costs of functional constipation: a population-based study. Int J Colorectal Dis 2011; 26:515-22.