Sindh Univ. Res. Jour. (Sci. Ser.) Vol: 40 (1) 45-48 (2008)

SURJ

Sindh Univ. Res. Jour.

A Statistical study of the Socio-economic factors causing Tuberculosis

Salahuddin, Irshad A., Arshad*, M. Karim* and Najma**

Department of Statistics, University of Peshawar, Peshawar. Pakistan

(Received 13th Jan. 2008 Revised 10th Feb.2008)

Abstract:

In this study an effort has been made to determine the socio-economic factors of Tuberculosis (TB) in Kohat District of Pakistan by using advanced statistical techniques. A random sample of 935 patients was selected for the purpose. The main factors discussed in this study were residence, age, sex, population and economic status. The best fitted log-linear model, consisting of three-factor interaction, four-factor interactions, and five factor interaction, was selected by means of backward elimination procedure. Binary Logistic regression was applied to interpret the interactions included in the model.

The interaction between pop (1) and eco (1) was significantly related to the log odds of TB positive. Tuberculosis, age, sex and population interaction was analyzed at different levels of age, sex, and population and the interaction among age (2), sex (1) and pop (1) found significantly related with log odds TB positive.

Keywords: Log-linear analysis, Backward Elimination Procedure, Binary Logistic regression

1. Introduction:

Tuberculosis, popularly known as TB is a chronic or acute bacterial infection that primarily attacks the lungs, but it may also affect any part of the body, particularly the kidneys, bones, lymph nodes, and brain. It is caused by Mycobacterium tubercle (a rod- shaped bacterium).

TB kills more people than any other single infectious disease throughout the world. The tuberculosis situation, like in many developing countries, is also very serious in Pakistan. It is one of the leading causes of death in developing countries. Half of all the untreated TB cases are fatal. It is principally a disease of poverty, with 95% of cases and 98% of deaths occurring in developing countries. TB is the biggest killer of women in the world. TB is present in almost every part of the world.

According to the World Health Organization (WHO); one individual becomes infected with TB every second and researchers have calculated that 8-10 million people contract the disease every year. Until the socio- economic factors that influence the spread of TB infection are remedied, there is no real possibility of completely eliminating the disease. Pakistan stands 6th in the world as far as TB are concerned. India, our neighbour is the first, with China holding the second position.

Spence *et al* (1993) obtained positive association between TB and poverty and even declared TB as determinant of poverty and poverty as determinant of TB. Gender tuberculosis research published by World Health Organization states that the incidence of pulmonary TB among women is always lower than in men. It says that nearly twice as many men have been died from TB as women in 2002. It also declares the age group 15-44 as the most common age of TB, the most important and productive age for both men and women.

Michael *et al* (2002) studied by means of multiple logistic regression analysis that housing density (average person per household), population and isolation type were positively

*Department of Mathematics and Statistics, Allama Iqbal Open University, Islamabad.
**Department of Statistics, Frontier Women University, Peshawar.

Scanned with CamScanner

associated with TB and average household income was negatively related with the occurrence of TB. The interaction of average person per household and average household income was significantly associated with TB.

2. Materials and Methods

A random sample of 935 Patients collected from various hospitals of Kohat district of Pakistan was investigated to determine the socio-economic factors of tuberculosis. Since all factors were categorical, therefore log-linear analysis was the appropriate technique to analyze the data. Log-linear modeling is essentially a discrete multivariate technique that is designed specially for analysis of data when both the independent and dependent variables are categorical or nominal. (Upton, 1978; Agresti, 1990).

Log-linear modeling has the ability to test relationships between multiple outcome and multiple explanatory variables. The use of simultaneous tests under the log-linear approach permits researchers to test the association between each outcome variable and each explanatory variable, as well as the associations between outcome variables for tests involving multiple outcome variables. (Fienberg, 1980; Upton, 1978; Agresti, 1990).

The simplest log-linear model for 2×2 contingency table is as follows:

$$\text{Ln}(F_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$$

$\text{Ln}(F_{ij})$ is the log of the expected cell frequency of the cases for cell ij in the Contingency table. μ denotes the overall mean of the natural log of the expected frequencies. Each λ represents "effects" which the variables have on the cell frequencies.

A and B stand for the variables. i and j refer to the categories within the variables. Therefore: λ_i^A is the main effect for variable A; λ_j^B stands for the main effect for variable B, and λ_{ij}^{AB} denotes the interaction effect for variables A and B. (Knoke and Broke, 1980).

3. Backward elimination procedure

The basic purpose of backward elimination is to find that unsaturated model that will provide the best fit to the data. It declares the highest order interaction as the saturated model with zero likelihood ratio chi-square and p-value equal to one. It then eliminates various insignificant interactions at various steps. The best fitted model is obtained at 20th iteration. It includes three factor, four factor interaction, and five factor interactions. Detail about the best fitted model is presented in the **Table 1.**

Table 1. Test of significance

If Deleted Simple Effect is	DF	L.R. Chisq. Change	Prob.	Iteration
TB × RES × AGE × SEX × ECO	2	4.98	0.08	4
RES × AGE ×ECO × POP	2	3.95	0.14	3
TB × AGE × SEX × POP	2	6.33	0.04	4
TB × ECO × POP	1	7.05	0.01	4

4. Binary Logistic Regression

It is a special type of regression in which the response variable is dichotomous and the predictor variables are discrete or a mixture of discrete and continuous. It is used to measure the instantaneous rate of change in probability of occurrence of an event with change in a given predictor. (Demeris, 1990).

Binary logistic regression model was applied to interpret those interactions which were included in the best log-linear model. In these models the variable of interest was considered as response and all the remaining variables included in that interactions were treated as explanatory variables. Binary logistic regression models were fitted for the detailed analysis of all those higher order interactions, which were included in the best model, by the help of Statistical Package for Social Sciences (SPSS).

5, Interaction between Tuberculosis, Economic status and Population.

The above model when studied by logistic regression shows that the interaction between pop (1) by eco (1) is significantly associated to the log odds of TB. The odds are significantly high, (1.947) times in favour of pop (1) by eco (1) interaction to lie in TB positive patients against TB negative patients.

Scanned with CamScanner

Table 2. Taking tuberculosis as response and population, economic status, and interaction between population and economic status as explanatory variables.

	B	SE	Wald	df	Sig.	Exp (B)	95.0%C1 for EXP(B) Lower	Upper
Set Pop (1)	.011	.183	.004	1	.952	.989	.691	1.415
1 ECO(1)	,315	.219	2.065	1	.151	1.370	.892	2.107
ECO(1)POP	.666	.296	5.064	1	.024	1.947	1.690	3.479
Constant	.472	.140	11.362	1	.001	1.602		

The said logistic regression model can also be expressed as:

Logit (P) = 0.472 - 0.011 pop (1) + 0.315 eco (1) + 0.666 pop (1) by eco (1).

Interaction among Tuberculosis, Age, Sex, and Population.

In the above model the interaction of age (2) by pop (1) by sex (1) is significantly associated with log odds of positive TB. The odds are significantly low, (0.417) times in favour of age (2) by pop (1) by sex (1) interaction to accommodate in the positive TB patients group compared to the negative TB patients group.

Table 3. Taking tuberculosis as response and age, sex, population, and interaction among age, sex, and population of a household as explanatory variables.

	B	S.E.	Wald	df	Sig.	Exp(B)	95.0%C.1. for EXP (B) Lower	Upper
Step 1. AGE			6.757	2	.034			
AGE(1)	-.364	.318	1.313	1	.252	.695	.373	1.295
AGE(2)	.244	.247	.975	1	.323	1.277	.786	2.073
SEX (1)	.454	.208	4.760	1	.029	1.574	1.047	2.367
POP (1)	.618	.193	10.259	1	.001	1.855	1.271	2.707
AGE POP SE			11.702	2	.003			
AGE(1) by POP by SEX(1)	.212	.565	.141	1	.708	1.236	.408	3.743
AGE (2)by POP by SEX (1)	-.875	.282	9.598	1	.002	.417	.240	.725
Constant	.233	.272	.737	1	.390	1.263		

The above logistic regression model can also be written as:

Log it (P) = 0.233 – 0.364 age (1) + 0.244 age (2) + 0.454 sex (1) + 0.618 pop(1) + 0. 212 age (1) by pop (1) by sex (1) – 0. 875 age (2) by pop (1) by sex (1).

Interaction among Tuberculosis, Residence, Age, Sex, and Economic Status.

It is an independent model because interaction between residence, age, sex and economic status are independent of the log odds of TB positive. Further residence at level (1). i.e. urban residence is significantly related to the log odds of positive TB, the variable of interest in the analysis. The odds are significantly high, (2.028) times that economic status at level (1). i.e. poor people will affect the TB contacted patients against patients without TB.

Scanned with CamScanner

Table 4. Assuming tuberculosis as response, residence, age, sex, and economic status as explanatory variables

	B	S.E.	Wald	df	Sig.	Exp (B)	95.0%C1 for EXP (B) Lower	Upper
Step 1. RES(1)	-.458	.145	9.965	1	.002	.633	.476	.841
AGE			2.979	2	.225			
AGE(1)	.286	.298	.915	1	.339	.752	.419	1.349
AGE (2)	.083	.238	.123	1	.726	1.087	.682	1.734
SEX(1)	.130	.245	.805	1	.370	1.139	.857	1.513
ECO (1)	.707	.148	22.909	1	.000	2.028	1.518	2.709
Constant	.564	.244	5.336	1	.021	1.759		

The above logistic regression model may also be expressed as:

Logit(P) = 0.564 – 0. 458 res (1) - 0.286 age (1) + 0.083 age (2) + 0.130 sex(1) + 0.707 eco(1)

6. Conclusion:

The main objective of this study was to determine the most probable risk factors of Tuberculosis in Kohat district of Pakistan. A random sample of 935 patients was collected for the completion of the above task. The response variable in this study was tuberculosis, a binary variable taking the value "2" for TB patients and "1" otherwise. The variables selected as predictors were residence, sex, age, population and economic status of household. Since all variables were categorical and the collected data was count data, therefore, log-linear analysis was applied to find possible risk factors of TB. Backward elimination procedure was then used to determine the most suitable parsimonious model for TB.

Tuberculosis, economic status and population interaction was the first three factor interaction in the best log-linear model. This interaction when tested at different levels of economic status and population by logistic regression showed significant association between the log-odds of positive TB and interaction between poor economic status and overcrowded population.

Tuberculosis, age and sex, and population interaction was the first four factor interaction of selected log-linear model. This interaction when analyzed at different levels of age, sex, and population by applying logistic regression showed that the interaction among age (2), pop (1) and sex (1). That is, overcrowded population of male sex of age 15-54 years was significantly related to the log odds of TB positive.

Tuberculosis, residence and age, sex and economic status interaction was the only five factor interaction of the best selected log-linear model. Fit of the model shows that the interaction among residence × age × sex × economic status was insignificantly related with the log odds of TB positive.

7. References:

Agresti A. (1990) "Categorical data analysis." New York: John Wiley.

Demaris, A. (1990) "Interpreting logistic regression results." A critical commentary. Journal of Marriage and the Family, vol: 52, 271-277

Fienberg. (1980) "The analysis of cross-classified categorical data." 2nd ed. Bostan, MA: MIT Press.

Knoke, D. and P. J. Burke (1980) "Log-Linear Models". Sage Publications, Inc. Newberry Park, California. USA.

Michael C, R. Peter E. Nowgesic (2002) "The of housing density, isolation and tuberculosis in Canadian First National Communities." International Journal of epidemiology, vol: 31, No. 5, 940-945.

Spence DPS, Hotchkiss J, Williams CDS, and Davies PDO. (1993)" Tuberculosis and Poverty." BMJ. **307**: 759-761.

Upton GJG. (1978) "The analysis of Cross-tabulated Data." John Wiley and Sons. Inc. New York, USA.

WHO "Department of Gender, Women and Health." Pdf Aug: (2003).

Scanned with CamScanner